Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Hybrid CNN-Transformer for Knee Osteoarthritis Severity Grading

Unnati Patel[1], Sanskruti Patel[2*], Dharmendra Patel[3], Niky Jain[4], Suchita Patel[5], Ronesh Gangavani[6]

[1,2,3,6]Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology, CHARUSAT Campus, Anand, Gujarat, India.
[4,5]Institute of Science & Technology for Advanced Studies & Research, CVM University, Vallabh Vidyanagar, Anand, Gujarat, India.

E-mail: [1]unnatipatel.mca@charusat.ac.in, [2*]sanskrutipatel.mca@charusat.ac.in, [3]dharmendrapatel.mca@charusat.ac.in, [4]niky.jain@cvmu.edu.in, [5]suchita.patel@cvmu.edu.in, [6]roneshgangavani.mca@charusat.ac.in

Orcid ID: [2*]0000-0002-1440-0668

Abstract

The problem of accurately diagnosing of the degree of severity of knee osteoarthritis (KOA) based on simple radiographic images lies in the difficulty of distinguishing between adjacent degrees and differences in imaging conditions, nad in the ordinal nature of the Kellgren-Lawrence (KL) scoring system. In this paper, we use plain radiography (X-ray anteroposterior knee radiographs) as the main type of imaging for KOA analysis. To solve these problems, we introduce an innovative hybrid architecture based on CNNs and transformers with adaptive feature integration and ordinal-aware dual-head learning for KOA degree of severity diagnosis. The novel architecture incorporates a CBAM-ResNeXt-50 model as the backbone network for texture extraction, along with a lightweight transformer-based encoder for modeling the whole anatomy structure. We effectively integrate local and global semantics by designing a learnable adaptive feature fusion module at the image level, producing stage-aware attention on different KOA degrees. Furthermore, we develop an ordinal-aware dual-head learning paradigm that can jointly conduct KL grade classification and continuous KOA severity regression tasks. Experimental results achieve 96.84% accuracy, 0.96 macro F1-score, 0.21 MAE, and 0.959 macro AUC-ROC with fewer adjacent-grade confusions.

Keywords: Computer-Aided Diagnosis, Deep Learning, Hybrid CNN-Transformer Architecture, Knee Osteoarthritis, KL Grading, Medical Image Analysis, X-ray Radiography.

1. Introduction

Knee osteoarthritis is a degenerative musculoskeletal condition that has an adverse impact on movement, quality of life, and the capacity to lead an independent lifestyle, especially among elderly patients. There has been an increase in the prevalence of knee osteoarthritis, owing to factors like population aging and mechanical strain caused by physical activities, thus posing a significant burden on public health care systems across the world. Radiography has continued to be the gold standard for the diagnosis of KOA due to its high availability, affordability, and suitability for general use. The most common scoring system used to evaluate the severity of KOA is the Kellgren-Lawrence (KL) classification system. This

*** Corresponding Author**

DOI: https://doi.org/10.36548/jiip.2026.2.010
Received: 05.03.2026, received in revised form: 06.04.2026, accepted: 25.04.2026, published: 09.05.2026

involves classifying the disease process in terms of five sequential grades, KL 0-4, using distinctive radiographic features including osteophytes, joint space narrowing (JSN), bony deformities, and subchondral sclerosis. However, despite having distinct criteria, there is always inconsistency in grading KOA using KL scoring. CNNs perform well in capturing local discriminative texture features from the pixels [1], [5], [6]. Nevertheless, due to limited receptive fields, CNNs cannot adequately capture any kind of spatial dependency that is required to understand the global anatomy. Hence, relevant contextual details about the knee joint often get missed by CNN models. This limitation generally leads to incorrect classification of borderline KL grades, particularly in cases where there is a subtle change in the structure [6], [12], [26], [28].

As a result of their use of self-attention mechanisms [11], [36], and [37], transformers are capable of understanding context on a larger scale, hence their popularity among researchers working in the area of medical imaging. This type of machine learning is able to recognize sophisticated joint relationships in cases of knee osteoarthritis [12], [13]. Unfortunately, these algorithms need large volumes of data and are vulnerable to noise and alterations in imaging conditions [11], [27].

This development resulted in the emergence of hybrid models based on combining the abilities of transformers to perform global reasoning with those of CNNs to extract local features. Despite the progress achieved by some newly developed KOA-specific hybrid models [39], there are still many issues to be resolved. The majority of current models utilize simple or static fusion techniques, which are not flexible enough to allow for any adaptation of the influence of local and global knowledge. Additionally, the ordinal nature of KL classification was overlooked by most previous approaches, which viewed this process as a completely independent one.

In this research, a novel deep learning architecture based on CNN and transformer with an ordinal-aware dual head approach and learnable adaptive fusion will be proposed to address the aforementioned challenges. The CNN backbone consists of the transformer encoder that tracks long-range structural dependencies within the knee joint along with the local texture feature extraction module (e.g., geometrical warping and mixed order phases). The model uses the learnable fusion coefficients to combine the two distinct feature domains for estimating the appropriate weights of each branch. The KL-grade classification and severity regression tasks will be performed concurrently to avoid contradictory predictions in terms of disease severity.

The use of the Tibial Spiking Knee OA Dataset provided by Mendeley Data [24], which is an openly available KL graded radiograph dataset and provides training, validation, and test sets based on KL grade 0 to 4, is a substantial strength. This dataset has been developed across a broad spectrum of radiographic phenotypes ranging from morphological changes related to tibial spiking to the common KOA symptoms, and it forms the basis for developing and testing deep learning models. In order to preprocess the radiographs uniformly and minimize variations resulting from different imaging techniques or anatomy, a standard preprocessing pipeline is employed.

A large body of research clearly proves that the architecture under consideration consistently surpasses the transformer-based architecture [36, 37], novel CNN architectures [35], [38], conventional CNN baseline architectures [31–34], and hybrid networks designed specifically for the KOA dataset [39, 40]. The necessity of the loss function for ordinal consistency and the adaptive fusion approach is substantiated by ablation study findings. Moreover, the Grad-CAM visualization technique confirms the potential practical significance

and interpretability of our model by demonstrating that it indeed concentrates on important clinical areas.

The key contributions of this research can be highlighted as follows:

- Proposed a hybrid CNN Transformer model to improve the accuracy of knee osteoarthritis severity evaluation from X-rays through the integration of global and local information.
- Proposed a learnable adaptive combination and ordinal-aware dual-head model to perform KL-grading using continuous severity regression while achieving an optimal balance between local and global data.
- Extensive experimentation and interpretability analysis of the proposed model through Grad-CAM visualization and ablation studies on public KOA datasets.

2. Literature Review

CNNs can directly learn hierarchical radiographic representations from raw image data, they have been essential to computer-aided KOA grading. It has been demonstrated that transfer learning improves performance even more when there are initially no labeled datasets. Training networks from scratch is not as effective as fine-tuning pre-trained VGG16 and ResNet, as shown by Serir et al. [5]. The contributions of Tiulpin and Saarakkala [6] developed a deep CNN that performed extremely well and was designed specifically for KL-gradation radiography. The architecture of EfficientNet enabled the sensitivity to small changes in the radiographs to be improved using compound scaling. Using efficient preprocessing algorithms, Zhang et al. [7] have shown that EfficientNet-B3 and EfficientNet-B7 perform better than earlier CNNs. Furthermore, Kumar and Goswami [1] have demonstrated that two different types of preprocessing methods could significantly improve early KOA detection.

However, CNN-based models have some limitations owing to their local receptive field structure. The connections longitudinally along the tibiofemoral joint, which are important for diagnosing the intermediate KL categories, are poorly represented. CNN-based models often find it challenging to distinguish between KL-1 and KL-2 as well as KL-2 and KL-3 when there is no clear evidence of radiographic changes, as seen in several studies [6], [12], [26], and [28]. Such limitations highlight the importance of developing models that can combine global structure and local cues.

Manually crafted descriptors with deep representations represent one of the hybrid solutions proposed to solve these limitations. Hybrid approaches attempt to integrate priors in the radiography domain that could potentially be used to boost the accuracy and interpretation of the KOA grade. For instance, Khalid et al. [2] introduced a hybrid method for KOA grading via analysis of X-rays that utilizes CNNs and manually engineered descriptors. To better distinguish between intermediate KL grades, better, Almusa et al. [8] introduced a solution that uses CNN features and texture descriptors, such as GLCM. To make the system invariant to environment-related variations (lighting conditions, contrast), Swapna et al. [9] proposed hybridization using SIFT-LBP and deep descriptors. Clinically relevant markers, such as joint space width (JSW), have been integrated into hybrid architectures by some previous studies. To achieve a higher ordinal consistency for KOA grading, Alavanthar et al. [4], [10] employed

JSW estimates in a multitask learning framework. Moreover, ensembling-based hybrids, that combine multiple outputs from different deep architectures were investigated [3].

Using hand-crafted features in these approaches is unjustified even when the model outperforms the basic CNNs. The inclusion of such features depends on the parameters and often fails to generalize to another dataset with varied resolution, noise, and image acquisition techniques. As such, deep hand-crafted fusions have failed to exhibit robustness in different radiography contexts despite their theoretical strength.

The transformer network can take advantage of self-attention mechanisms to capture long-term dependencies. This has become an essential feature of transformers that makes them a suitable choice for medical image analysis. This is particularly important in assessing KOA since it is always essential to make connections between spatial relationships across the joint to distinguish the level of disease severity.

Chen et al. [11] introduced the TransUNet, a significant development that improves semantic segmentation accuracy by combining transformer encoders with CNN architectures. Based on this, researchers have developed transformer-based graders that concentrate on KOA. Sekhri et al. [13] proposed a Swin Transformer-based approach for the extraction of hierarchical anatomical representations, while Antony et al. [12] incorporated both channel and spatial attention mechanisms to address the ambiguous borderline KL class. In addition, vision transformers (ViT-B/16 [36], Swin Transformer-T [37]), which are general-purpose transformers, have been utilized for KL grading, outperforming conventional CNNs in most cases. The importance of fusing convolutional knowledge with global context understanding is further illustrated by KOA-specific hybrid transformers, such as the latest [39].

Transformers are highly sensitive to the presence of anatomical variations, noise, and poor image quality, which are inherent features of real-life clinical images. Furthermore, the application of transformers in numerous medical fields remains restricted due to their dependence on large-scale labeled data.

Because of differences in acquisition methods, patient positioning, radiation doses, and scanner types, radiographic images utilized in KOA assessment exhibit high variability. Pre-processing pipelines are thus needed to enhance the discrimination of KL grades and model generalizability. As reported by Mohammed et al. [14], denoising and normalization of radiographs enhance predictive accuracy. The widely applied CLAHE algorithm has proven effective in boosting joint space contrast without creating any artifacts [18]. ROI segmentation has likewise demonstrated notable advantages. As observed by Malik et al. [15], knee joint localization is facilitated by grade differentiation and lowers variability. Many works [16], [17], [20], and [21] have indicated that class imbalance, especially the scarcity of KL-4 samples, poses a challenge and that majority class dominance could be mitigated using oversampling.

Harmonization methods have been explored for the purpose of minimizing within-dataset variance. In cases where the model needs to generalize under contrast variations and non-canonical anatomy, methods such as normalization, registration based on shape, and augmentation under controlled conditions [14], [18], [22], and [30] increase consistency. This is particularly critical due to the presence of multiple morphologies in the Tibial Spiking KOA dataset utilized in this study [24].

CNN-only frameworks have drawbacks. Adjacent KL grades are frequently misclassified due to CNNs' inability to identify global anatomical links (as reported in [5-7], [26], [28]).

There are certain difficulties with deep handcrafted hybrid models. As shown in [8], [9], these models exhibit poor generalization performance, inadequate scalability, and fragility under various imaging settings.

There is a challenge for KOA models built on transformer-based systems. KL labels represent a grading scale as discussed [11–13], [39], the overwhelming number of the existing models relies on transformers, requiring huge amounts of input data, being sensitive to images' distortions, and not paying any attention to ordinal-aware learning. Most recent hybrid CNN Transformers for image segmentation fail to include ordinal-aware learning, although they offer better performance in terms of interpretation and context awareness [19].

There is an issue with the pipelines for fragmented preprocessing. As mentioned above, preprocessing algorithms are usually used independently of each other and are excluded from the hybrid model architectures, affecting the efficiency of the pipeline negatively. Table 1 demonstrates that the accuracy of previous studies using CNN and transformer-based models varies between 88% and 95%

Table 1. Summary of Performance Metrics Reported in Previous KOA Classification Studies

Model	Accuracy (%)	Macro F1	MAE	RMSE
VGG16 [31]	88.72	0.84	0.52	0.71
ResNet50 [32]	90.15	0.86	0.47	0.66
DenseNet121 [33]	91.04	0.87	0.44	0.63
EfficientNet-B3 [7], [34]	92.56	0.89	0.40	0.59
EfficientNetV2-S [35]	93.48	0.90	0.37	0.55
ConvNeXt-S [38]	94.02	0.91	0.34	0.52
ViT-B/16 [36]	94.65	0.92	0.32	0.50
Swin Transformer-T [37]	95.12	0.93	0.29	0.47
KOA Hybrid Transformer (2023) [39]	95.38	0.93	0.27	0.45
Selective Shuffled Transformer (2023) [39]	95.71	0.94	0.26	0.44

The design of the CNN Transformer model, integrating adaptive feature fusion, sign-consistent dual loss learning, and standard pre-processing, developed within this study to provide a more precise, robust, and clinically meaningful grading system for KOA under these constraints, is driven by these limitations collectively.

3. Materials and Methods

3.1 Dataset Description

Training, developing, validating, and testing of the KOA classifiers were all carried out using publicly accessible data. The images employed in this study have been sourced from the "Tibial Spiking Knee OA Dataset" provided by Mendeley Data [24]. It consists of anteroposterior knee radiographic images classified under the Kellgren Lawrence (KL) grading system. Each of the five classes of KL grading (0 to 4), starting from healthy joint structure through to highly osteoarthritic joints, is well represented within this dataset. Train, validation, and test datasets came from the Tibial Spiking Knee OA Dataset [24]. Furthermore, an independent database provided on Kaggle was employed to build the auto_test subset in order

to test the generalizability of the proposed model. Knee radiographs with KL scores (0-4) are contained in this dataset, which was solely used for robustness assessment purposes. This practice facilitates cross-dataset evaluation and reduces the likelihood of overfitting [40]. Anteroposterior knee radiographs or plain radiographic X-ray imaging constitute the primary imaging modality employed for assessing KOA in this paper.

Sample radiographic images of the anteriorposterior view of the knee from the dataset, demonstrating the variation in stiffness levels for KL grade classification (0 to 4), not divided into regions, are depicted in Figure 1.

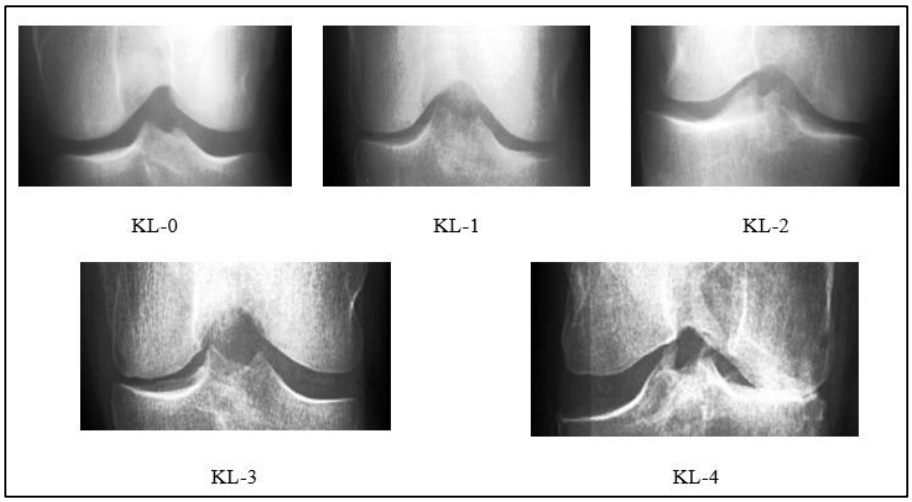

Figure 1. Sample Knee Radiographs across KL Grades (0 - 4)

The distribution of samples across splits and KL grades is summarized in Table 2.

Table 2. Distribution of KL Grades in the Primary Mendeley Dataset and External Auto_Test Set

KL Grade	Train	Validation	Test	Auto_test	Total
KL-0	2286	328	639	604	3857
KL-1	1046	153	296	275	1770
KL-2	1516	212	447	403	2578
KL-3	757	106	223	200	1286
KL-4	173	27	51	44	295
Total	5778	826	1656	1526	9786

3.2 Class Imbalance Considerations

There is an intrinsic class imbalance issue in the dataset, with lower KL grades (0–2) being substantially higher than higher grades (3–4). To overcome this limitation, only the training set was subjected to dynamic augmentation and class-aware sampling, which overrepresented minority class images without subtly altering the underlying data structure. To maintain the integrity of performance evaluation and ensure that all published results reflect genuine generalization, the validation, test, and auto_test splits are not enhanced. Even though KL-4 contains less data (295 images), evaluation fairness was preserved by reporting macro-averaged metrics (Macro-F1 and Macro-AUC), which evaluate each class equally, independent of sample frequency. During training, class-aware sampling and targeted augmentation were also employed to reduce bias against minority classes. In order to ensure that each mini-batch contained a more equal distribution of KL grades and prevent bias toward majority classes, minority classes like KL-4 were managed during training using class-aware sampling and targeted augmentation.

3.3 Preprocessing Overview

The appearance of knee radiographs varies greatly depending on the scanner hardware, exposure settings, patient position, soft-tissue composition, and architecture. Clinically significant signals, particularly those related to joint-space width, osteophyte formation, and subchondral bone texture, may be obscured by this type of variability, which may also result in unstable or unreliable feature learning. We used a structured, standardized pipeline to preprocess every image to lessen these issues. Each pipeline component sought to improve radiographic clarity, account for nuisance variation, and reveal morphological characteristics that are essential for KL grading. Even though the original radiographs are grayscale, each image is duplicated over three channels to satisfy the input requirements of pretrained CNN architectures.

3.3.1 Intensity Normalization

To lessen variability based on exposure and illumination, pixel values were standardized to a specific range [a,b] (usually [0,1]). The formula in (1) is used to determine the normalized intensity value for each pixel p.

$$I_{norm}(p) = a + (b - a)\,\frac{I(p) - I_{min}}{I_{max} - I_{min}} \tag{1}$$

Here, $I(p)$ is the intensity of the image at pixel p. Here, $I_{\min} = \min_{p\in\Omega} I(p)$ and $I_{\max} = \max_{p\in\Omega} I(p)$ are minimum and maximum intensity over the image domain Ω, respectively. The parameters a and b determine the intensity range to be reached, typically [0,1], usually From now on we denote by the normalized intensity at pixel p is denoted by $I_{\text{norm}}(p)$. This scaling enhances numerical stability and ensures contrast-related features are consistently represented.

3.3.2 Contrast-Limited Adaptive Histogram Equalization (CLAHE)

CLAHE enhances local contrast while preventing noise over-amplification by operating on small contextual tiles. CLAHE was used in this investigation with a tile grid size of 8 x 8 and a clip limit of 2.0, which effectively enhanced contrast without over-amplifying radiography noise. The histogram is calculated for every tile T using the formula in (2):

$$h(i) = \sum_{p\in\mathcal{T}} 1(I(p) = i) \tag{2}$$

Here, $h(i)$ is the total number of pixels in T with intensity level i, where $i \in \{0, \dots, L-1\}$ and L are the gray of levels. Where p is a pixel coordinate, its intensity $I(p)$ and the indicator function $\mathbb{1}(\cdot)$.

By creating a histogram for each image tile, CLAHE increases local contrast. It produces an improved and noise-controlled image with an adjusted histogram by capping the histogram at a predetermined threshold, clipping the surplus, and distributing it uniformly over all gray levels rather than allowing noise to run wild. From that updated histogram, it then extracts a cumulative distribution function that instructs you on how to translate the original pixel intensity to its enhanced value. It employs bilinear interpolation to maintain seamless transitions between adjacent tiles without creating border artifacts. In general, this technique

reduces noise while improving local contrast. The entire procedure can be summed up as follows (3):

$$I_{\text{CLAHE}}(p) = \sum_{k=1}^{4} w_k \ (L-1) \frac{\sum_{j=0}^{I(p)} h_{\text{adj},k}(j)}{|T_k|} \tag{3}$$

where $h_{\text{adj},k}$ is the clipped and redistributed histogram of the k-th neighboring tile, and $w_k$are interpolation weights satisfying $\sum_{k=1}^{4} w_k = 1$.

3.3.3 Region of Interest (ROI) Extraction

A bounding box was used to isolate the tibiofemoral joint in order to guarantee attention to diagnostically significant features. The region of interest (ROI) is extracted from the original image, as mentioned in (4):

$$I_{\text{ROI}}(u, v) = I(x_0 + u, y_0 + v), \tag{4}$$

In this case, I(x, y) represents the original image in global coordinates, whereas $I_{\text{ROI}}(u, v)$ indicates the extracted region of interest stated in local coordinates (u, v). The top-left corner of the bounding box utilized for ROI extraction is defined by the parameters x_0 and y_0.

When anatomical landmarks $L_i = (x(L_i), y(L_i))$ were available, the bounding box origin is computed as defined in (5):

$$x_0 = \min_i \ x(L_i) - \delta_x, y_0 = \min_i \ y(L_i) - \delta_y \tag{5}$$

The i-th anatomical landmark on the knee joint is indicated by $L_i = (x(L_i), y(L_i))$, where $x(L_i)$ and $y(L_i)$ are its horizontal and vertical coordinates. The ROI borders are extended beyond the extreme landmark positions using the specified margin parameters δ_x and δ_y. The ROI bounding box's width and height are then calculated using the formula in (6):

$$w = \max_i \ x(L_i) - x_0 + 2\delta_x, h = \max_i \ y(L_i) - y_0 + 2\delta_y \tag{6}$$

In this case, w and h stand for the bounding box's width and height that surround the anatomical region of interest. To maintain anatomical consistency between samples, these dimensions are calculated from the maximal landmark coordinates with respect to the bounding box origin, with symmetric margin extensions managed by δ_x and δ_y.

ROI extraction removes irrelevant structures and enforces anatomical alignment across images.

3.3.4 Bilateral Filtering

To suppress high-frequency noise while preserving joint boundaries, bilateral filtering was applied. As stated in (7), the filtered intensity at pixel p is calculated as follows:

$$I_{\text{bil}}(p) = \frac{1}{W(p)} \sum_{q \in \mathcal{N}(p)} G_{\sigma_s}(\| p - q \|) \, G_{\sigma_r}(| I(q) - I(p) |) \, I(q) \tag{7}$$

In this case, q specifies an adjacent pixel within the spatial neighborhood $\mathcal{N}(p)$, and $I_{\text{bil}}(p)$ indicates the filtered intensity at pixel p. The initial intensity at pixel q is denoted by the word $I(q)$. The functions $G_{\sigma_s}(\cdot)$ and $G_{\sigma_r}(\cdot)$ denote the spatial and range Gaussian kernels, parameterized by σ_s and σ_r, which weight contributions based on spatial distance $\| p - q \|$ and intensity difference $| I(q) - I(p) |$, respectively, enabling edge-preserving smoothing with normalization. The computation of the normalization factor W(p) is provided in (8):

$$W(p) = \sum_{q \in \mathcal{N}(p)} G_{\sigma_s}(\| p - q \|) \, G_{\sigma_r}(| I(q) - I(p) |). \tag{8}$$

In this case, $W(p)$ stands for the normalization factor, which is calculated by adding the spatial and range kernel responses over the neighborhood $\mathcal{N}(p)$. This normalization guards against bias resulting from local kernel weighting and guarantees appropriate scaling of the filtered output. The spatial kernel G_{σ_s} preserves locality, while the range kernel G_{σ_r} preserves edges, improving clarity of osteophyte borders and joint margins.

3.3.5 Image Resizing

After rescaling the images for consistency with architectural considerations and uniform processing of receptive fields, all ROI extracted images underwent scaling by bilinear interpolation.

In combination, these preprocessing steps enhance anatomical accuracy, level out radiographic information across individuals, and facilitate the accurate detection of important features. Figure 2 presents the complete procedure for preparing knee radiographs.

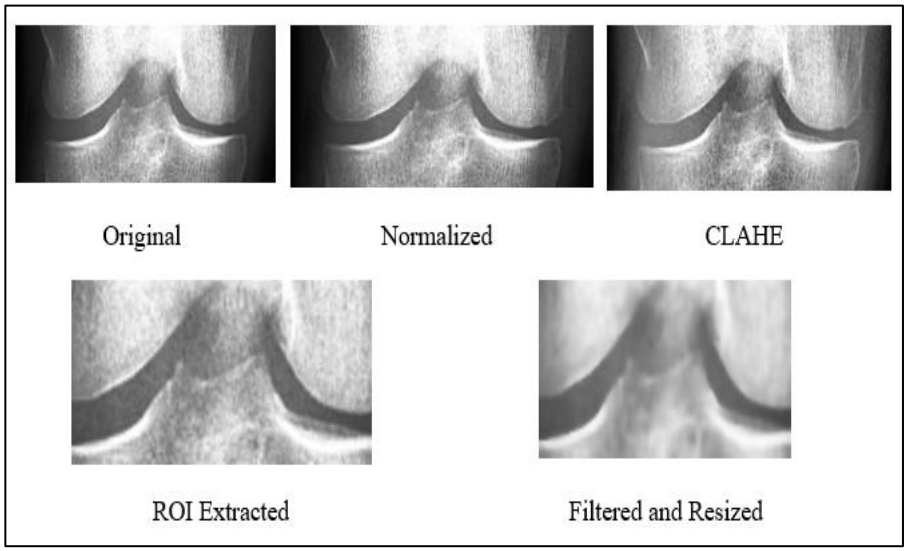

Figure 2. Preprocessing Pipeline Visualization

4. Methodology

The proposed framework incorporates a hybrid approach to local feature extraction using CNNs, along with global reasoning using transformers along with the integration of a preprocessing approach that considers boundaries, adaptive data augmentation, and class balancing techniques. The architecture of the proposed hybrid CNN transformer model for

automatic grading of KOA based on knee X-ray images is illustrated in Figure 3 below. Four main parts make up the model: (i) an ordinal-aware dual-head prediction module; (ii) parallel CNN and transformer-based feature extraction; (iii) adaptive feature fusion; and (iv) a standardized preprocessing pipeline. Each element is intended to tackle particular issues related to radiographic KOA evaluation, such as the ordinal nature of KL grading, global anatomical variability, and inter-grade visual similarity.

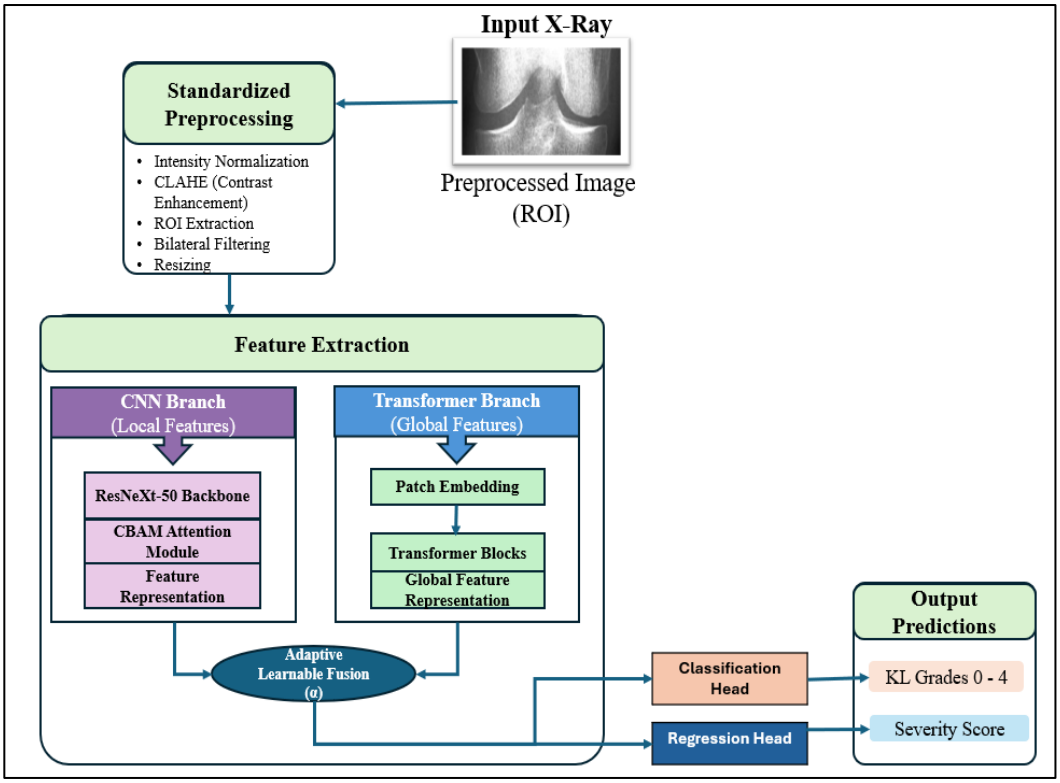

Figure 3. Detailed Architecture of the Proposed Hybrid CNN Transformer Model for Knee Osteoarthritis Severity Grading

4.1 Preprocessing Pipeline

As described in Section 3.3, preprocessed radiographs were used for all studies. Images were scaled and normalized after preprocessing before being fed into the model. To enable a fair and consistent comparison, all baseline models and the architecture mentioned above were trained under the same set of conditions. In this challenge setup, the pre-defined train, val, test, and auto_test splits were already provided by the Mendeley dataset.

4.2 A Proposed Hybrid CNN Transformer Architecture

We suggest a hybrid architecture that combines a lightweight transformer encoder for global contextual modeling with CBAM-enhanced ResNeXt-50 for local feature extraction. This approach is founded on the clinical understanding of KOA evaluation, where grading can benefit from both large-scale structural patterns (like overall joint alignment and compartmental constriction) and fine-resolution textural signals (like osteophyte growth and trabecular alterations). An overview of the suggested hybrid CNN Transformer architecture is shown in Figure 3. It includes the dual-head prediction framework, learnable feature fusion component, a transformer encoder for global context modeling, and CBAM-boosted ResNeXt-based local feature-learning.

4.2.1 Local Feature Extraction via ResNeXt-50 with CBAM

To improve the representational cardinality and aid the network in learning fine-grained radiographic textures, ResNeXt-50 [29] makes use of grouped convolutions. This increases the network's computational efficiency. A Convolution Block Attention Module (CBAM) was added to the respective values in the final three stages of the backbone to increase the discriminative ability of these features. To highlight diagnostically significant patterns (such as small osteophyte borders and changes in trabecular density) and down weight uninformative areas, CBAM performs channel and spatial attention sequentially.

A. ResNeXt Grouped Convolution Block

The ResNeXt block's grouped convolution operation is specified as follows in (9):

$$U = \text{Concat}_{k=1}^{K}(W_k * X) \tag{9}$$

In this case, W_k is the convolution kernel of the k-th group, and X is the input feature map to the ResNeXt block. Convolution is indicated by the operator $*$, and the cardinality—that is, the number of grouped convolution paths—is specified by the operator K. The aggregated feature representation U is created by concatenating the outputs of each group along the channel dimension. Next, the residual output of the ResNeXt block is calculated using the formula in (10):

$$F_{\text{res}} = X + \phi\left(W_{\text{proj}} * U\right) \tag{10}$$

The output feature map of the ResNeXt block following residual fusion is indicated here by the symbol F_{res}. The activation function $\phi(\cdot)$, which is implemented as ReLU, comes after the projection convolution W_{proj} is applied to the aggregated features U to guarantee dimensional compatibility. Effective gradient propagation is made possible by the residual link, which adds the modified features to the input X.

Figure 3 presents the detailed dual-path architecture of the proposed hybrid CNN Transformer model, including feature extraction, adaptive fusion, and ordinal-aware dual-head prediction. Stacking blocks and a final global mapping yield the CNN feature representation as defined in (11).

$$F_{CNN} = ConvBlock_{head}(X) \in \mathbb{R}^{H_c \times W_c \times C_c} \tag{11}$$

In this case, the input feature map derived from the stacked convolutional backbone layers is represented by X. The final convolutional head that combines hierarchical features to create a compact representation is represented by the operator $\text{ConvBlock}_{\text{head}}(\cdot)$. The final feature tensor, $F_{\text{CNN}} \in \mathbb{R}^{H_c \times W_c \times C_c}$, where C_c is the number of output channels that capture high-level semantic information and H_c and W_c are the spatial dimensions.

The resulting feature map is denoted as F_{CNN}.

B. CBAM - Channel Attention

The attention map as described in (12) is calculated by the channel attention mechanism:

$$M_C(F) = \sigma\left(W_2\,\mathrm{ReLU}\big(W_1 f_{\mathrm{avg}}\big) + W_2\,\mathrm{ReLU}(W_1 f_{\mathrm{max}})\right) \tag{12}$$

Here, f_{avg} and f_{max} are the channel-wise global average and maximum pooled descriptors, respectively, and F stands for the input feature map. The channel attention map $M_C(F)$ is created by processing these features through shared fully connected layers W_1 and W_2, ReLU, and sigmoid activation $\sigma(\cdot)$. As stated in (13), the refined feature representation is obtained:

$$F' = M_C(F) \odot F \tag{13}$$

Here, F'represents the channel-refined feature map obtained via element-wise multiplication, which adaptively enhances informative channels while suppressing irrelevant responses.

C. CBAM - Spatial Attention

As stated in (14), the spatial attention mechanism calculates the refined CNN feature representation and the spatial attention map.

$$M_S(F') = \sigma\left(\mathrm{Conv}_{7\times7}([f_c^{\mathrm{avg}}; f_c^{\mathrm{max}}])\right),$$
$$F_{\mathrm{CNN}} = M_S(F') \odot F', \tag{14}$$

In this case, the spatial feature maps generated by channel-wise average and max pooling are denoted by f_c^{avg} and f_c^{max} respectively, while channel-wise concatenation is denoted by [;] The spatial attention map M_S is created by processing the concatenated features using a 7×7 convolution $\mathrm{Conv}_{7\times7}$. This map is then applied to the CNN feature representation F_{CNN} in order to highlight important spatial areas.

4.2.2 Global Feature Extraction via Transformer Encoder

A scaled-down transformer is used to model global dependencies in the joint domain. Each windowed patch in the contaminated ROI is mapped to an embedding vector. Spatial linkages are maintained and positional embeddings can be learned. Even though transformers often require massive datasets, the proposed method uses a lightweight transformer encoder with just two layers together with a strong CNN backbone that extracts robust local representations. This hybrid solution reduces the quantity of data required while enabling global contextual modeling.

A. Patch Embedding

Each image patch's embedding is calculated using the formula given in (15):

$$z_i^0 = E\,\mathrm{vec}(P_i) + p_i \tag{15}$$

In order to balance computational speed and spatial detail, a 16 × 16 patch size was selected, yielding 196 tokens for a 224 × 224 input. The transformer encoder models global contextual dependencies using two layers with four attention heads. In this case, E is a learnable linear projection matrix, $\mathrm{vec}(\cdot)$ is the vectorization operator, and P_i represents the i-th picture

patch. The initial token embedding z_i^0 input to the transformer encoder is obtained by adding the positional encoding p_i to preserve spatial information. For an input token sequence

$$Z = [z_1, z_2, \dots, z_N]$$

The transformer encoder applies multi-head self-attention to model global contextual patterns such as compartmental asymmetry and alignment shifts.

B. Transformer Encoder Layer

According to (16), the transformer encoder layer uses feed-forward transformations and multi-head self-attention to update the token representations:

$$\tilde{z}_i^l = z_i^{l-1} + \mathrm{MHSA}(\mathrm{LN}(z^{l-1}))_i$$
$$z_i^l = \tilde{z}_i^l + \mathrm{MLP}(\mathrm{LN}(\tilde{z}^l))_i \quad (16)$$

The representation of the i-th token at transformer layer l is indicated here by z_i^l. Layer normalization is indicated by the operator $\mathrm{LN}(\cdot)$, while the multi-head self-attention module and position-wise feed-forward network are represented, respectively, by MHSA and MLP. The intermediate token representation following the attention-based residual update is denoted by the word $\tilde{z}_i^l$. Given the moderate dataset size typical of medical imaging research, a shallow transformer architecture with two encoder layers was chosen to minimize computational complexity and prevent overfitting.

C. Multi-Head Self-Attention (MHSA)

The transformer encoder's multi-head self-attention operation is calculated using the formula given in (17):

$$\mathrm{MHSA}(Q,K,V) = \mathrm{Concat}_{h=1}^{H}\left(\mathrm{softmax}\left(\frac{QW_Q^{(h)}(KW_K^{(h)})^\top}{\sqrt{d_k}}\right)VW_V^{(h)}\right)W_O \quad (17)$$

In this case, the query, key, and value matrices are denoted by Q, K, and V. Learnable projections for the h-th attention head are represented by the matrices $W_Q^{(h)}$, $W_K^{(h)}$, and $W_V^{(h)}$, where $h = 1, \dots, H$ and H indicate the number of heads. The output projection over the concatenated head outputs is carried out by the matrix W_O, and the key dimensionality utilized for attention scaling is indicated by d_k.

4.3 Adaptive Learnable Fusion Mechanism

Most hybrid models use static weights or fixed concatenation to combine CNN and transformer features. However, each KL grade has varied diagnostic performance for local versus global features. Early-stage KOA relies on minor textures, while advanced-stage KOA necessitates global structure analysis. Because the fusion coefficient α was set to 0.5 at the beginning of training, CNN and transformer features contributed equally. The parameter was then optimized during training via backpropagation.

To address this, a learnable fusion coefficient $\alpha \in [0,1]$ was introduced, and the fused representation is computed as defined in (18):

$$F=\alpha F_{CNN}+(1-\alpha)F_{TR} \quad (18)$$

In this case, F_{CNN} stands for the local characteristics that the CNN extracted, and F_{TR} stands for the global features that the transformer generated. The fused feature representation F is produced by controlling the relative contribution of the two representations using the learnable weight $\alpha \in [0,1]$.

The fusion coefficient α is trained by backpropagation, allowing the network to learn to adaptively balance local and global feature representations according to image properties.

4.4 Dual-Head Prediction Module

While KOA severity is ordinal, most deep learning techniques treat KL grading as nominal. We created a dual-head prediction module to clearly depict the ordered relationship between various disease progression stages.

4.4.1 Classification Head

The head of the classifier takes global average pooling and is extended by a fully connected layer with Softmax activation that yields the final discrete class probabilities $\hat{y}_{\text{class}}$ as defined in (19).

$$s = \text{FC}_{\text{cls}}\big(\text{GAP}(F)\big), \hat{y}_{\text{class}} = \text{softmax}(s) \quad (19)$$

The global average pooled feature vector from the fused representation F is indicated here by $\text{GAP}(F)$. The class logits generated by the classification layer FC_{cls}, are represented by the vector s, and the predicted class probabilities $\hat{y}_{\text{class}}$ are acquired using softmax normalization.

4.4.2 Regression Head

Parallel to classification, a regression head predicts a continuous severity score $\hat{y}_{\text{reg}}$, as defined in (20) capturing ordinal proximity between KL grades.

$$\hat{y}_{\text{reg}} = \text{FC}_2\left(\text{ReLU}\left(\text{FC}_1\big(\text{GAP}(F)\big)\right)\right) \quad (20)$$

In this case, the predicted continuous severity score produced by a regression head made up of two fully connected layers, FC_1 and FC_2, is represented as $\hat{y}_{\text{reg}}$. The final output is obtained by passing the global average pooled feature $\text{GAP}(F)$ through FC_1 with ReLU activation and then mapping it by FC_2.

4.4.3 Ordinal-Aware Dual Loss

A. Cross-Entropy Loss

The cross-entropy function, as stated in (21), is used to calculate the classification loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log\left(\hat{y}_{\text{class},c}\right) \quad (21)$$

In this case, the ground-truth indicator for class c is denoted by y_c, and the predicted probability for class c is represented by $\hat{y}_{\text{class},c}$. The five Kellgren-Lawrence (KL) grades are represented by the parameter C, which represents the total number of classes.

B. Smooth L1 Loss (Huber Variant)

The Smooth L1 (Huber) function, which is defined in (22), is used to calculate the regression loss:

$$\mathcal{L}_{\text{SmoothL1}} = \begin{cases} \frac{1}{2}(t - \hat{y}_{\text{reg}})^2/\delta, & | t - \hat{y}_{\text{reg}} | < \delta, \\ | t - \hat{y}_{\text{reg}} | - \frac{1}{2}\delta, & \text{otherwise,} \end{cases} \tag{22}$$

In this case, the ground-truth ordinal severity label is represented by $t \in \{0, \dots, 4\}$, and the projected continuous regression score is represented by $\hat{y}_{\text{reg}}$. By defining the transition point between the quadratic and linear regimes of the loss, the parameter δ —typically set to 1— improves robustness to outliers. The smooth-L1 loss was selected because it promotes ordinal consistency through continuous severity regression and offers stable optimization and robustness to outliers.

C. Total Dual Loss

The classification and regression losses as stated in (23) are combined in the overall training objective.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{SmoothL1}} \tag{23}$$

In this case, L is the final joint loss results from merging the regression and classification losses and λ is a learnable weight that regulates the regression objective's contribution.

The dynamic weighting of these tasks ensures that they are mutually beneficial to accomplish, alleviating the misclassification between adjacent KL grades.

The model was trained for a maximum of 60 epochs with a batch size of 32 and usually stopped early due to convergence before reaching the maximum number of epochs. Hyperparameters were chosen in a way that is fair to compare with the baseline models.

5. Experiments

The experimental configuration of this study was designed to rigorously evaluate the suggested CNN Transformer architecture in both within-domain and cross-domain scenarios. A multi-metric performance evaluation, including classification accuracy, ordinal consistency, and interpretability, has been taken into consideration; extensive experiments have been conducted and compared with classical models and state-of-the-art baseline models; ablation analyses have been presented to assess the contributions of individual architectural components. To further examine the trained model's capacity for generalization, radiographs from each KL grade were included in an external auto_test dataset collected from Kaggle [40].

5.1 Computational Environment

All the experiments were conducted on a high-end workstation equipped with an NVIDIA RTX series graphics card with 16GB VRAM and an Intel Core (10th generation) CPU with 32GB of system memory. The software environment consisted of Python 3.10 and PyTorch 2.11, which was the main deep learning framework. For the preprocessing of images, OpenCV was used, while Albumentations was used for augmentation laundering, considering that it is fast in executing radiographic transformations. Graphical representation and statistical analysis were done using Matplotlib and Seaborn. The flexibility to integrate convolutional and transformer modules was seamless and easy in PyTorch. It could conduct mixed precision computations and was efficient in memory management. At an average inference speed of about 20ms/image on an NVIDIA RTX graphics card, the proposed hybrid CNN-Transformer model has approximately 29.6 million trainable parameters. To ensure accurate and repeatable depiction of experimental data, all graphs and visualizations in this work were created using Microsoft Excel and Python libraries (Matplotlib and Seaborn). Table 3 provides a high-level summary of the suggested hybrid CNN–Transformer architecture, including its main processing steps and matching output dimensions.

Table 3. High-Level Architecture Summary of the Proposed Model

Stage	Description	Output Size
Input	Preprocessed Image	224 x 224 x 3
CNN Branch	ResNeXt-50 with CBAM attention module	2048
Transformer	Patch embedding followed by transformer encoder	256
Fusion	Adaptive Fusion	256
Output	Classification head and regression head	-

5.2 Baseline Architectures for Comparative Evaluation

To obtain a sound and equitable evaluation, the proposed networks were evaluated against a diverse range of baseline architectures.

5.2.1 Classical and Modern CNN Baselines

Various popular CNN architectures used for radiographic image analysis were re-implemented as basic models:

- VGG16 [31]: It is a classical deep convolutional network, which has stable feature representations and has been applied in clinical imaging tasks for years.
- ResNet50 [32]: A residual network that includes identity skip connections to address the vanishing gradient problem, which helps in training deep models better.
- DenseNet121 [33]: A network using dense connections that aim to alleviate the vanishing-gradient problem and encourage the flow of gradients, with many reused features between layers.
- EfficientNet-B3 [7], [34]: A compound-scaled network that uses a combination of depth, width, and resolution to achieve good accuracy and efficiency trade-offs.
- EfficientNetV2-S [35]: A CNN based on the next generation of more quickly converging, better performing architecture for both medical and natural images.

These state-of-the-art CNN baselines enable a clear evaluation of the benefits of using transformer-based global contextual reasoning in the hybrid architecture.

5.2.2 Transformer and Hybrid Architecture Baselines

We included the following transformer-based and hybrid models to benchmark our proposed model against other state-of-the-art architectures for global contextual reasoning:

- Vision Transformer (ViT-B/16) [36]: A competitive instance-based global attention model for variable sequence lengths due to its strong ability to model long-range dependencies.
- Swin Transformer-T [37]: A hierarchical transformer with shifted windows that has been successfully applied in the fields of medical images and natural images.
- ConvNeXt-S [38]: A state-of-the-art convolutional architecture designed based on transformer-style construction principles, offering significant performance improvements in CNNs.

KOA-specific hybrid transformer models re-implemented from [39] (CNN transformer hybrids and attention-augmented KL grading systems built for osteoarthritis evaluation).

5.3 Hyperparameter Configuration

The eventual hyperparameters were selected after a controlled grid search process, enriched by knowledge of former deep learning research in radiographic KOA evaluation. Table 4 shows all settings applied during training, such as the optimizer configuration and learning rate schedule, the batch size and loss-balancing parameters. Grid search was used to determine the learning rate over the range $\{1\times10^{-5}, 3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}\}$, with 1×10^{-4} producing the best validation result.

Table 4. Hyperparameter Settings Used for Training All Models

Hyperparameter	Value / Setting
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
LR Scheduler	Cosine Annealing
Weight Decay	0.01
Batch Size	32
Number of Epochs	60 (with early stopping)
Dropout (classification head)	0.3
Dropout (transformer encoder)	0.1
Patch Size (Transformer)	16 × 16
Embedding Dimension (Transformer)	256
Number of Attention Heads	4
Fusion Weight Initialization (α)	0.5
Loss-Balancing Coefficient (λ)	Learnable (initialized at 0.5)
Activation Function	ReLU (CNN branch), GELU (Transformer)
Image Input Size	224 × 224 × 3

5.4 Evaluation Metrics

Multi-class classification performance was evaluated using a set of widely recognized metrics.

5.4.1 Accuracy

Classification accuracy is calculated using the formula in (24):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{24}$$

where TP, TN, FP, and FNdenote true positives, true negatives, false positives, and false negatives, respectively.

5.4.2 Precision, Recall, and F1 Score

For each class k, precision and recall were defined as expressed in (25):

$$\text{Precision}_k = \frac{TP_k}{TP_k+FP_k}, \qquad \text{Recall}_k = \frac{TP_k}{TP_k+FN_k} \tag{25}$$

where the numbers of true positives, false positives, and false negatives for class k are represented by TP_k, FP_k, and FN_k, respectively. The class-wise F1 score was computed accordingly. Macro-averaging was used to reduce bias from class imbalance.

5.4.3 Cohen's Kappa

To capture inter-class agreement beyond chance levels, Cohen's Kappa statistic was computed as defined in (26):

$$\kappa = \frac{p_o-p_e}{1-p_e} \tag{26}$$

where $p_o$is the observed agreement and $p_e$is the expected agreement by chance [25].

5.4.4 AUC for Multi-Class ROC

Class-specific ROC curves were computed using a one-vs-rest approach and their mean was reported.

5.5 Regression Metrics for Ordinal Consistency

As the order of KL grades reflects disease severity, Kappa statistics are inadequate, and regression-based metrics are used for testing ordinal consistency.

5.5.1 Mean Absolute Error (MAE)

The Mean Absolute Error is an average of the absolute errors without considering their direction. It is defined as (27):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i | \tag{27}$$

where $y_i$denotes the true KL grade and $\hat{y}_i$represents the predicted continuous severity.

5.5.2 Root Mean Squared Error (RMSE)

RMSE is a lack of average squared prediction error measures and square root of penalty for large deviations. It is defined as (28):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{28}$$

where, again, $y_i$is the true KL grade and $\hat{y}_i$is the predicted severity score.

6. Results

The suggested hybrid architecture is thoroughly evaluated in this section, which includes quantitative comparisons against baselines, class-wise diagnostic performance, ordinal regression analysis, the impact of preprocessing, ablation studies to gauge the significance of each architectural component, and qualitative insights from gradient-based interpretability techniques. Consistent performance across the internal test set and the external auto_test dataset proved the robustness of the proposed approach.

6.1 Quantitative Comparison with Baseline Models

The Hybrid CNN Transformer model showed superior performance to all the baseline models in terms of accuracy, macro F1-score, and ordinal regression measures on the test set of Mendeley. It obtained an accuracy of 96.84%, with a macro F1-score of 0.96 and a mean absolute error (MAE) value of 0.21; thus state-of-the-art performance was achieved relative to all compared models.

The system showed macro-Precision at 0.9391 and macro-Recall at 0.9329, indicating high classification abilities in all the KL groups. Cohen's Kappa is almost perfect at 0.9236. Macro-AUC-ROC is at 0.9592, which indicates a clear separation of KOA into its three forms: mild, moderate and severe. All these metrics demonstrate that the hybrid architecture proposed in this study is clinically meaningful, ordinal-consistent and stable.

CNNs (including VGG16, ResNet50, DenseNet121 and EfficientNet-B3) have yielded decent results; however, they did not manage to recognize subtle differences between morphological structures in the images corresponding to different stages of KOA. CNNs have performed quite well but are still inferior to the transformer-based architectures, mostly due to their inability to recognize morphologies.

Recognizing the global context is essential for KOA diagnostics, as evidenced by the success of the transformer baselines (ViT-B/16 and Swin Transformer-T), which yielded superior or comparable results compared to the CNNs. However, even the best-performing transformer baselines showed poorer results, especially concerning borderline cases, since they are not able to accurately analyze fine-grained radiographic textures.

KOA-specific hybrid transformers proposed in recent studies [39] yield better results. However, their application of fixed fusion strategies and the lack of ordinal-sensitive learning restrict their competitive performance with the proposed model.

As shown in Table 5, the proposed hybrid CNN Transformer model consistently outperforms conventional CNNs, transformers, and recent hybrid baselines, achieving the highest accuracy and macro F1-score while minimizing MAE and RMSE.

Table 5. Diagnostic Metrics for the Proposed Model

Metric	Value
Accuracy	96.84%
Macro F1	0.96
MAE	0.21
RMSE	0.37
Precision (Macro)	0.9391
Recall (Macro)	0.9329
Cohen's Kappa	0.9236
AUC-ROC (Macro)	0.9592

6.2 Class-Wise Performance and Confusion Patterns

Class-wise analysis revealed excellent discrimination for all KL grades. The model additionally performed extremely well for KL-0, KL-2, and KL-4 where morphological differences are more distinct. The rate of borderline misclassification was, however, significantly lower between KL-1 ↔ KL-2 and KL-2 ↔ KL-3, indicating that the ordinal-aware dual-head structure contributed toward apparently insignificant differences in disease severity.

The resultant confusion matrix displays substantially fewer near-grade errors than any of the CNN or transformer-based baselines, confirming the robustness and diagnostic accuracy of our model. Figure 4 shows the confusion matrix of our model on the test set, demonstrating good-class-wise discrimination across each KL grade and a low frequency of near-grade misclassifications. Misclassifications between KL-1 and KL-2 are substantially lower than CNN baselines, according to the confusion matrix study, suggesting enhanced sensitivity to minute structural variations.

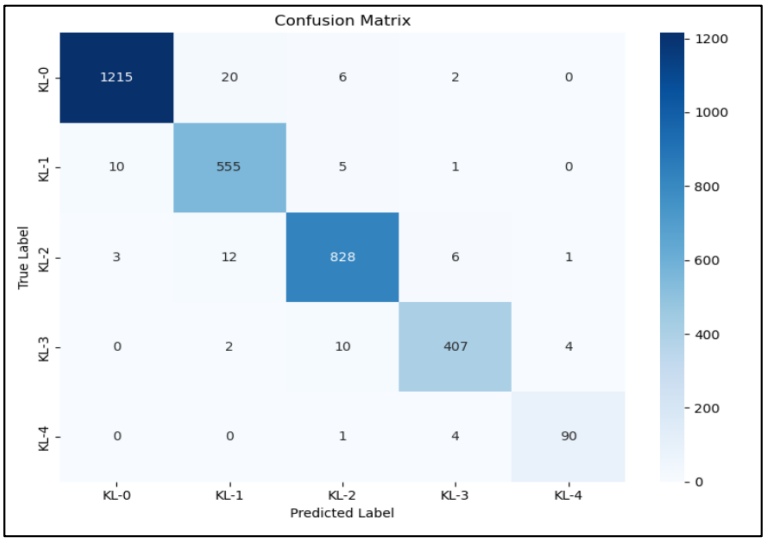

Figure 4. Confusion Matrix of the Proposed Model

6.3 Regression Behavior and Ordinal Consistency

With KL grade scores ranked by increasing levels of severity, evaluating the continuous output of the models in terms of severity also provides additional insights into their validity.

The models trained with the regression head disabled or with fixed loss weights had significantly worse consistency with the ordinal labels. Excluding the regression head, in particular, resulted in an MAE increase to 0.38, demonstrating that ordinal constraint learning helps significantly.

Results from evaluating the performance of our models in terms of regression are summarized in Table 6.

Table 6. Regression Performance for Ordinal KOA Severity Prediction

Metric	Value
Mean Absolute Error (MAE)	0.21
Root Mean Squared Error (RMSE)	0.37

6.4 Ablation Study

The role of each architectural component is estimated using ablative analysis. Performance was incrementally improved by incorporating increasingly stronger components, including (in order) CBAM, the transformer encoder, the fusion layer, and the dual loss function, starting from the ResNeXt-50 backbone.

Table 7 illustrates that performance increases consistently as the components CBAM, transformer blocks, adaptive fusion, and dual loss learning are sequentially incorporated.

Table 7. Ablation Study Demonstrating Contribution of Each Component

Configuration	Accuracy	F1-score	MAE
ResNeXt-50 only	92.15%	0.89	0.38
ResNeXt-50 with CBAM	93.42%	0.9	0.34
ResNeXt-50 with CBAM and Transformer	94.12%	0.92	0.31
Hybrid with Fixed Fusion	94.72%	0.93	0.29
Hybrid with Learnable Fusion	96.12%	0.95	0.24
Final Proposed Model	96.84%	0.96	0.21

6.5 Explainability Analysis

The Grad-CAM visualization results have highlighted that the activation is localized to the joint space region around the medial and lateral locations, tibial plateau, femoral condyles, osteophyte margins, and subchondral sclerosis zones. The proposed technique always emphasizes the relevant regions. The prediction emphasis of the network is continuously constrained to the clinically relevant anatomical regions related to osteoarthritis development, as can be observed from Figure 5, which shows sample Grad-CAM images across KL grades.

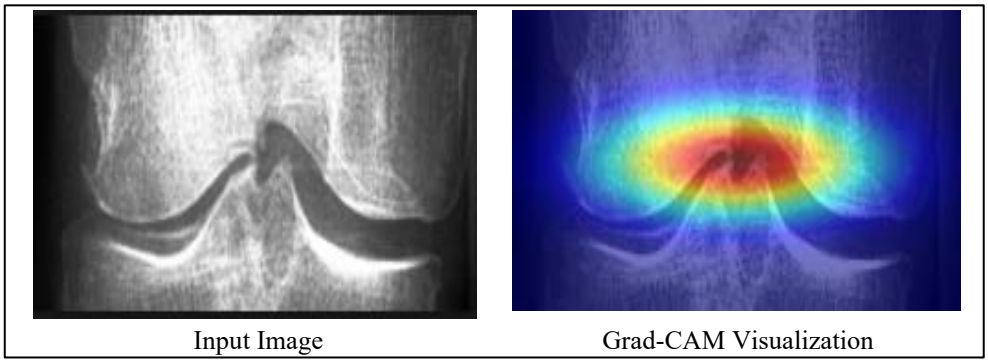

Figure 5. Grad-CAM Visualizations Showing Anatomical Regions Driving KOA Predictions Across KL Grades

This region is consistent with the biomarkers used by clinicians for KL scoring. This facilitates cooperation between humans and AI systems in clinical decision-making processes, similar to the findings of Tschandl et. al [23]. This is also an indication that our model is not a black box; it is interpretable in the clinical environment. Table 8 summarizes the attention score values obtained from the aggregation of all the regions.

Table 8. Average Grad-CAM Attention Scores for Anatomical Regions

Anatomical Region	**Attention Score (0-1)**	**Clinical Interpretation**
Medial Joint Space	0.87	Primary site of KOA progression
Lateral Joint Space	0.82	Important for symmetry assessment
Tibial Plateau	0.78	Indicates subchondral sclerosis
Femoral Condyles	0.74	Critical load-bearing region
Marginal Osteophyte Borders	0.85	Key marker of early-moderate KOA
Tibial Spine Area	0.63	Secondary morphological cue
Surrounding Soft Tissue	0.21	Low relevance - good specificity

7. Discussion

The proposed hybrid model achieves superior performance in the KOA task and outperforms CNNs, convolutional transformer models, recently proposed transformer-only models, and earlier released disease focused hybrid systems.

7.1 Role of Local and Global Representations in KOA Assessment

CNNs are excellent models for learning local textures; however, they have limitations such as a limited receptive field and an inability to capture cross-compartment relationships. The transformer models complement them by modeling long-range dependencies. However, they oversmooth fine-grained texture information or require a large amount of data for stable training. A consistent balance between the learned local information from CNNs and global information through transformers is established in each epoch, as reflected in the stability of the fusion coefficient α from the second epoch.

This proposed approach addresses these shortcomings through a combination of three components:

- ResNeXt-50 with CBAM, which focuses on clinically relevant local features.
- A Transformer Encoder with relatively low computational complexity that aggregates global structural information from the joint.
- A fusion module that dynamically controls the impact of the above two modules.

The attention to the extracted features can dynamically modify the focus based on the disease stage, such that texture information is emphasized during the early stages of KOA development, followed by structural cues in more advanced stages of the disease. The proposed fusion framework is consistent with clinical observations where grade 1 can be characterized using only textural features, whereas higher grades must include global deformity patterns. The trend of the learnable fusion parameter (α) in relation to the KL grade is illustrated in Figure 6.

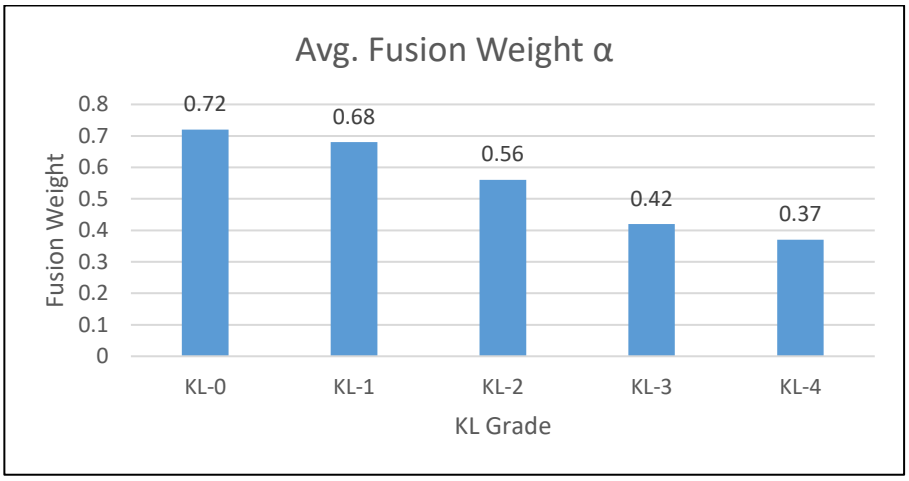

Figure 6. Average Fusion Weight (α) for Local vs. Global Features Across KL Grades

7.2 Limitations

Although our proposed framework has shown strong performance, it is still associated with several limitations. First, the model was developed and tested on a single publicly available dataset. The Mendeley dataset is heterogeneous but may not sufficiently represent the variability in actual clinical practice. External validation on independent cohorts—OAI, MOST, or institution-level datasets would be necessary to validate the generalizability of our models to other imaging protocols, demographic strata and acquisition settings.

Second, the dataset uses image-level KL grades that may not capture inter-observer variability between radiologists. Although the proposed model utilizes an ordinal learning process, it does not adequately consider diagnostic uncertainty or annotation disagreement. Future work may consider incorporating probabilistic or soft labels, multi-reader consensus scoring and/or ensemble-based uncertainty estimation to bring the model closer in line with the real-world diagnostic variability it is trained against.

Third, the transformer encoder reads 2D patches of a single anterior-posterior radiographic view. Thus, the model does not take advantage of richer anatomical evidence such as bilateral symmetry, 3D joint structure and longitudinal developments. Adding more views of radiographs, temporal follow-ups or geometry-aware representation can improve the diagnostic robustness.

Finally, hybrid architecture strikes a good balance of accuracy and efficiency, it is still more computationally exhaustive than typical CNNs. Considerations related to optimization for runtime, model compression, and deployment in low-power or mobile and point-of-care settings were outside the scope of our present study but require further exploration.

8. Conclusion

The proposed design integrates local texture learning and global context modeling to present a novel CNN-Transformer based model for automatic KOA severity grading. Compared to existing approaches like CNNs, transformers, and their hybrids, the proposed framework outperforms with 96.84% accuracy, a macro F1 score of 0.96, a mean absolute error (MAE) of 0.21, and a macro-AUC-ROC of 0.959. The proposed design exhibits excellent ordinal consistency and classification accuracy. The combination of adaptive fusion, CBAM-

based feature extraction, and ordinal-oriented dual-head prediction helps minimize misclassification cases among adjacent KL scores. Additionally, the suggested method has demonstrated its strong generalizability, both in internal and external datasets. As a result, the clinical applicability of the suggested approach is very promising. The present research suggests a novel framework that can be effectively used for computer-assisted assessment of KOA severity. Future work should focus on uncertainty modeling and optimizing clinical applicability.

References

[1] Goswami, Agam Das. "Automatic Classification of the Severity of Knee Osteoarthritis Using Enhanced Image Sharpening and CNN." Applied Sciences 13, no. 3 (2023): 1658.

[2] Khalid, Ahmed, Ebrahim Mohammed Senan, Khalil Al-Wagih, Mamoun Mohammad Ali Al-Azzam, and Ziad Mohammad Alkhraisha. "Hybrid Techniques of X-Ray Analysis to Predict Knee Osteoarthritis Grades Based on Fusion Features of CNN and Handcrafted." Diagnostics 13, no. 9 (2023): 1609.

[3] Pi, Sun-Woo, Byoung-Dai Lee, Mu Sook Lee, and Hae Jeong Lee. "Ensemble Deep-Learning Networks for Automated Osteoarthritis Grading in Knee X-ray Images." Scientific Reports 13, no. 1 (2023): 22887.

[4] Alavanthar, Logeshwari, Jayashree Stalin, and K. Jasmine Mystica. "Deep Learning-Based Framework for Automated Classification of Knee Osteoarthritis Severity and Detection of Joint Space Width in X-Ray Imaging." In International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024), Atlantis Press, 2025, 1152-1161.

[5] SERIR, Amina, Lynda Bounif, Rania Lounaci, and Yamina Mezerna. "Deep Learning Framework for Assessing Knee Osteoarthritis Severity." In 2025 2nd International conference on Advances in Electronics, Control and Communication Systems (ICAECCS), IEEE, 2025, 1-6.

[6] Tiulpin, Aleksei, and Simo Saarakkala. "Automatic Grading of Individual Knee Osteoarthritis Features in Plain Radiographs Using Deep Convolutional Neural Networks." Diagnostics 10, no. 11 (2020): 932.

[7] Pan, Jian, Yuangang Wu, Zhenchao Tang, Kaibo Sun, Mingyang Li, Jiayu Sun, Jiangang Liu, Jie Tian, and Bin Shen. "Automatic Knee Osteoarthritis Severity Grading Based on X-ray Images Using a Hierarchical Classification Method." Arthritis research & therapy 26, no. 1 (2024): 203.

[8] Almusa, Lubna Mohammad, Turky Nayef Alotaiby, Hanan Saeed Murayshid, and Rawad Awad Alqahtani. "Hybrid Ensemble Model for Knee Osteoarthritis Grading: Integrating CNNs with GLCM Features and XAI." Diagnostics 16, no. 4 (2026): 539.

[9] Swapna, Munnangi, Mohammad Omar Sabri, Sugunakar Mamidala, SNaveen Kumar, and Chandi Priya KG. "Leveraging Deep Learning Methodology to Automate Knee Osteoarthritis Identification based on X-ray Images." In 2025 International Conference on Recent Innovation in Science Engineering and Technology (ICRISET), IEEE, 2025, 1-9.

[10] VR, Gokul Thamp, and T. Anjali. "Deep Learning and XAI for Knee Osteoarthritis Detection on X-Rays." In 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, 2025, 1925-1931.

[11] Chen, Jieneng, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. "Transunet: Transformers Make Strong Encoders for Medical Image Segmentation." arXiv preprint arXiv:2102.04306 (2021).

[12] Antony, Joseph, Kevin McGuinness, Noel E. O'Connor, and Kieran Moran. "Quantifying Radiographic Knee Osteoarthritis Severity Using Deep Convolutional Neural Networks." In 2016 23rd international conference on pattern recognition (ICPR), IEEE, 2016, 1195-1200.

[13] Sekhri, Aymen, Mohamed A. Kerkouri, Aladine Chetouani, Marouane Tliba, Yassine Nasser, Rachid Jennane, and Alessandro Bruno. "Automatic Diagnosis of Knee Osteoarthritis Severity Using Swin Transformer." In Proceedings of the 20th International Conference on Content-Based Multimedia Indexing, 2023, 41-47.

[14] Mohammed, Abdul Sami, Ahmed Abul Hasanaath, Ghazanfar Latif, and Abul Bashar. "Knee Osteoarthritis Detection and Severity Classification Using Residual Neural Networks on Preprocessed X-ray Images." Diagnostics 13, no. 8 (2023): 1380.

[15] Malik, Sanjeev, and Nikita Singhal. "A Comparative Analysis of Deep Learning Approaches for Knee Osteoarthritis Detection Using Indian and Multi-Centric Datasets." In 2025 International Conference on Ambient Intelligence in Health Care (ICAIHC), IEEE, 2025, 1-6.

[16] Srivastava, Sameer, Eshanee Ghosh, Abhinav Kumar, Parthiv Chahar, Arpit Utkarsh, and Raghavendra Mishra. "Multi-Class Deep Learning Architecture for COVID-19, Tuberculosis, and Pneumonia Classification Using Chest X-ray Images." Journal of Medical Imaging and Radiation Sciences 56, no. 6 (2025): 102115.

[17] Qu W, Balki I, Mendez M, Valen J, Levman J, Tyrrell PN. Assessing and Mitigating the Effects of Class Imbalance in Machine Learning with Application to X-ray Imaging. Int J Comput Assist Radiol Surg. 2020 Dec;15(12):2041-2048. doi: 10.1007/s11548-020-02260-6. Epub 2020 Sep 23. PMID: 32965624.

[18] Momenpour, Thomures, and Arafat Abu Mallouh. "Optimizing CNN-Based Diagnosis of Knee Osteoarthritis: Enhancing Model Accuracy with CleanLab Relabeling." Diagnostics 15, no. 11 (2025): 1332.

[19] Djoumessi, Kerol, Samuel Ofosu Mensah, and Philipp Berens. "A Hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Disease Detection from Retinal Fundus Images." In International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, Cham: Springer Nature Switzerland, 2025, 106-116.

[20] Wang, Jason, and Luis Perez. "The Effectiveness of Data Augmentation in Image Classification Using Deep Learning." Convolutional Neural Networks Vis. Recognit 11, no. 2017 (2017): 1-8.

[21] Tliba, Marouane, Yassine Nasser, Mohamed Amine Kerkouri, Aladine Chetoauni, and Rachid Jennane. "A Graph-Driven Approach to Knee Osteoarthritis Severity Classification." In 2025 33rd European Signal Processing Conference (EUSIPCO), IEEE, 2025, 1592-1596.

[22] Kuriyama, Yuya, Mitsuhiro Nakamura, and Megumi Nakao. "Data Augmentation Using the Hierarchical Encoding of Deformation Fields Between CT Images." IEEE Transactions on Radiation and Plasma Medical Sciences 8, no. 8 (2024): 939-949.

[23] Tschandl, Philipp, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda et al. "Human–Computer Collaboration for Skin Cancer Recognition." Nature medicine 26, no. 8 (2020): 1229-1234.

[24] Moilanen, M., Grönholm, T., Paloneva, J., & Äyrämö, S. (2024). Tibial Spiking Knee OA Dataset (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/6gbptmgp3y.1.

[25] McHugh, Mary L. "Interrater Reliability: The Kappa Statistic." Biochemia medica 22, no. 3 (2012): 276-282.

[26] Tajbakhsh, Nima, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?." IEEE transactions on medical imaging 35, no. 5 (2016): 1299-1312.

[27] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All You Need." Advances in neural information processing systems 30 (2017).

[28] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778.

[29] Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 1492-1500.

[30] Usama, Mohd, Emma Nyman, Ulf Näslund, and Christer Grönlund. "A Domain Adaptation Model for Carotid Ultrasound: Image Harmonization, Noise Reduction, and Impact on Cardiovascular Risk Markers." Computers in Biology and Medicine 190 (2025): 110030.

[31] Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv preprint arXiv:1409.1556 (2014).

[32] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778.

[33] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 4700-4708.

[34] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks." In International conference on machine learning, PMLR, 2019, 6105-6114.

[35] Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller Models and Faster Training." In International conference on machine learning, PMLR, 2021, 10096-10106.

[36] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv preprint arXiv:2010.11929 (2020).

[37] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." In Proceedings of the IEEE/CVF international conference on computer vision, 2021, 10012-10022.

[38] Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A Convnet for the 2020s." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 11976-11986.

[39] Sekhri, Aymen, Mohamed A. Kerkouri, Aladine Chetouani, Marouane Tliba, Yassine Nasser, Rachid Jennane, and Alessandro Bruno. "Automatic Diagnosis of Knee Osteoarthritis Severity Using Swin Transformer." In Proceedings of the 20th International Conference on Content-Based Multimedia Indexing, 2023, 41-47.

[40] Islam, N. (2023). Knee Osteoarthritis Grad-CAM [Computer Software]. Kaggle. https://www.kaggle.com/code/naim99/knee-osteoarthritis-grad-cam