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

A Hierarchical Swin Transformer-Based Framework for Knee Osteoarthritis Severity Classification Using Radiographic Images

Zahoor M. Aydam[1*], Baidaa Mutasher Rashed[2], Mohanad A. Kadhim[3]

[1]Department of Information Technology, College of Computer Science and Mathematics, University of Thi-Qar, Thi-Qar, Iraq.
[2]Department of Computers, Faculty of Computer Science and Mathematics, University of Thi-Qar, Nasiriyah, Thi-Qar, Iraq.
[3]Department of Artificial Intelligence, College of Science, University of Alkafeel, Najaf, Iraq.

E-mail: [1*]zahoor.mosad@utq.edu.iq, [2]baidaaalsafy@utq.edu.iq, [3]mohanad.assim@alkafeel.edu.iq

Orcid ID: [1*]0000-0001-9967-204X, [2]0000-0002-5193-0334, [3]0009-0009-1869-8084

Abstract

Knee osteoarthritis (KOA), one of the leading causes of musculoskeletal disease worldwide, results in severe pain, decreased physical activity, and poor quality of life particularly in the elderly population. Early evaluation of KOA staging is critical for treatment planning and decision-making for patients and clinicians. Currently, KOA grading mostly depends on subjective diagnosis based on orthopedic experience with radiograph analysis, which is inefficient and prone to inter-observer variability. In this work, a hierarchical deep learning framework based on Swin Transformer is presented for an automated KOA classification task from knee radiographs. The framework leverages hierarchical shifted-window self-attention to extract image features from local and global perspectives. It also utilizes a weighted random sampling strategy, a class-weighted cross-entropy loss function, and label smoothing to alleviate the class imbalance problem. The experiment is performed on a publicly available dataset with 8,260 knee radiographs containing five Kellgren-Lawrence (KL) grading categories: healthy, doubtful, minimal, moderate and severe knee osteoarthritis. In the experiment, a patient-wise dataset partitioning strategy is applied to address possible data leakage issues and provide a more reasonable assessment. The results showed that the proposed framework achieved 93.4% accuracy, 92.8% precision, 92.5% recall, and a 92.6% F1-score. Grad-CAM analysis indicates that the proposed framework attends to key KOA-related anatomical structures, which supports interpretation. Although the performance on the evaluated dataset shows it can achieve a competitive result, validation on multicenter datasets from different sources is still needed to fully demonstrate its generalization capability and robustness.

Keywords: Knee Osteoarthritis, Swin Transformer, Deep Learning, Medical Image Classification, Radiographic Grading, Class Imbalance and Hierarchical Vision Transformer.

*** Corresponding Author**

DOI: https://doi.org/10.36548/jiip.2026.3.025
Received: 28.05.2026, received in revised form: 30.06.2026, accepted: 22.07.2026, published: 05.08.2026

1. Introduction

Osteoarthritis (OA) is a progressive and chronic disease that specifically affects articular cartilages and underlying bone, resulting in chronic pain, joint stiffness, and eventual progressive loss of joint function. The knee joint is one of the commonly affected joints; when affected, it is commonly known as knee Osteoarthritis (KOA), which contributes a significant health burden globally. About 365 million people worldwide suffer from KOA, a number that is expected to increase due to the aging population [1]. The socioeconomic burden includes healthcare costs and other costs related to reduced productivity, decreased quality of life etc. The Kellgren-Lawrence (KL) classification, based on radiographic features is one of the most commonly adopted methods for quantifying the severity of KOA [2]. KL grading contains five grades, from Healthy (grade 0) to Severe (grade 4). The grades are defined by the presence of a specific combination of osteophytes, joint space narrowing, sclerosis, and deformities in X-ray images. However, this method may be subjective in nature.

With the rapid advancement of deep learning methods, a new approach has emerged to interpret medical images. It enables the development of AI algorithms capable of achieving expert-level diagnostic performance in several medical imaging tasks by analyzing images for the diagnosis of numerous pathologies [3, 4]. Convolutional Neural Networks (CNNs) have dominated the field by virtue of their ability to leverage hierarchical feature learning for creating highly informative representations from medical images [3, 4]. However, CNNs remain limited by local receptive fields.

As noted by Dosovitskiy et al. [5], Vision Transformers (ViT) revealed that transformer models could accurately capture global context in images with self-attention mechanisms. Nevertheless, conventional ViT requires massive training datasets and powerful computing power, along with quadratic computational complexity with respect to image resolution, making them inefficient in handling large-scale medical imaging data. In contrast, Liu et al. [6] proposed a new hierarchical structure based on the shifted window self-attention approach in the Swin Transformer to enable linear computational complexity while preserving model performance to extract local and global features.

The research presented herein extends the above strengths and proposes a hierarchical deep learning model using Swin Transformer as the backbone for KOA grading using knee radiographs. The contributions of this work include

1. Using Swin Transformer as a reliable hierarchical backbone for multi-scale feature extraction.

2. Designing a task-specific classification pipeline optimized for five-class KOA severity grading.

3. Establishing a balanced training dataset to solve the issue of class imbalance using weighted sampling, along with class-weighted cross-entropy loss in addition to the label smoothing technique.

4. Integrating established optimization and regularization techniques within a unified Swin Transformer-based framework for KOA grading, utilizing the AdamW optimizer, cosine annealing, and mixed- precision training.

The remainder of this paper is organized as follows: Section 2 discusses relevant literature, Section 3 describes the proposed theory along with its mathematical formulation, Section 4 outlines the adopted methodology, Section 5 reports and discusses findings, and Section 6 concludes and points out directions for future work.

2. Related Work

Automatic detection of knee osteoarthritis through the use of radiographic images is an area that has attracted many scholarly studies in recent years. Initially, these approaches involved handcrafted features such as texture features, HoG, and LBP combined with classic machine learning models. These approaches, however, showed poor results as they required domain knowledge to engineer features.

2.1 CNN-Based Approaches

Since the advent of deep learning, Tiulpin et al. [7] suggested a Siamese CNN model that predicted KOA grades based on the simultaneous analysis of bilateral knee X-rays. The model achieved a Cohen's kappa coefficient of 0.83, demonstrating the potential of transferring knowledge from large natural image databases. Sekhri et al. [8] proposed a Swin Transformer-based framework for automatic knee osteoarthritis severity assessment, demonstrating the effectiveness of hierarchical transformer architectures in capturing both local anatomical structures and global contextual information from radiographic images, Antony et al. [9] applied fine-tuned VGGNet and AlexNet models pre-trained on the ImageNet database for automatic classification of KOA using the Kellgren-Lawrence scale, which showed an improved accuracy of 9-15%.

Gorriz et al. [10] proposed an attention-based CNN framework for KOA severity assessment and explored strategies to improve minority-class representation. Thomas et al. [11] studied the use of ensemble techniques that utilize several CNN architectures (ResNet-50, DenseNet-121, InceptionV3) through majority voting, performing favorably on the OAI database. A deep learning method that incorporated ordinal loss for automatic KOA grade classification was recently presented by Chen et al. [12] This method achieved better results than existing methods in distinguishing between adjacent KL grades.

2.2 Transformer-Based Approaches

The positive outcomes of using transformers in NLP have drawn researchers to explore similar transformer architectures in medical imaging. Shamshad et al. [13] conducted a survey of the use of transformers in medical image analysis and noted their strengths compared to traditional CNNs in terms of long-range dependency modeling. Nguyen et al. [14] applied the ViT model to grading KOA along with data augmentation and transfer learning methods, reporting an accuracy of 88.2%. However, a substantial decline in minority classes was found due to class imbalance issues.

Badawy et al. [15] proposed a hybrid CNN-Transformer model where ResNet-50 was used as the backbone to extract local features, and subsequent transformer layers modeled global context to achieve an accuracy of 92.7%. This model also provided greater robustness to radiographic image artifacts. In particular, Sekhri et al. [16] argued for the importance of the

shifted-window attention mechanism of the Swin Transformer architecture in identifying discriminative local and global radiographic features for assessing KOA severity.

However, while there has been increasing popularity in using transformer architectures in medical imaging, studies specifically employing a Swin Transformer-based framework to address KOA grading with simultaneous solutions to class imbalance and optimization techniques are relatively few. Table 1 provides a comparative summary of related works on automated KOA classification. It should be noted that as different datasets and evaluation schemes were used, these results are solely intended for qualitative comparison.

Table 1. Comparative summary of related works on automated KOA classification

Reference	Method	Dataset	Validation Strategy	Classes	Accuracy
Tiulpin et al. [7]	Siamese CNN	OAI	Image-wise test set	5	66.7%
Sekhri et al. [8]	Swin Transformer	OAI + MOST	Train/Validation/Test split	5	70.17%
Antony et al. [9]	Fine-tuned VGGNet	MOST	Image-wise	5	60.3%
Gorriz et al. [10]	CNN with Attention	OAI + MOST	Train/Validation/Test split	5	71.2%
Thomas et al. [11]	Deep CNN	OAI	Image-wise	5	78.5%
Panwar et al. [14]	Vision Transformer	OAI	Image-wise	5	88.2%
Patel et al. [15]	Hybrid CNN–Transformer	Private	Train/Validation/Test split	5	92.7%
Proposed Model	Hierarchical Swin Transformer	OAI	Patient-wise	5	93.4%

3. Theoretical Framework and Mathematical Modeling

The mathematical formulation follows the conventional Swin Transformer as in [6], thus detailed derivations are omitted to save space. The mathematical concepts that underlie the proposed model are described in this section. The main characteristics of the Swin Transformer model and the custom designed classifier head are:

3.1 Swin Transformer Architecture

The Swin Transformer (also known as Shifted Window Transformer) is an image transformer using a hierarchical approach that enables self-attention computation within non-overlapping local windows but provides interaction between different windows through the concept of shifted windows. The model uses four levels of feature extraction processes, which progressively reduce spatial resolution while increasing channel dimensionality [13].

3.1.1 Patch Embedding

Let $I \in R^{H*W*C}$ be the input knee X-ray image, where H, W, and C are the height, width, and number of channels, respectively. The Swin Transformer starts by dividing the image into non-overlapping patches of size $P \times P$. In our current model, $\mathrm{P} = 4$ and $\mathrm{C} = 3$ (RGB), which gives us $\mathrm{N} = \frac{\mathrm{H}}{\mathrm{P}} \times \frac{\mathrm{W}}{\mathrm{P}}$ patches. Each patch is then flattened and embedded in a D-dimensional space let $x_p^i \in R^{P^2.C}$ be the raw pixel values of the i-th patch. The patch embedding is expressed as:

$$z_0 = [\, x_p^1 E;\ x_p^2 E; \ldots;\ x_p^N E] + E_{pos} \quad (1)$$

Where $E \in R^{P^2.C.D}$ is the learnable patch projection matrix, $E_{pos} \in R^{N.D}$ is the positional embedding, and $N = \frac{H}{P} \times \frac{W}{P}$ [5][6].

3.1.2 Window-based Multi-head Self-Attention (W-MSA)

The conventional multi-head self-attention layer suffers from quadratic computation cost regarding the number of tokens, which makes its application in high-resolution images computationally expensive. This issue is solved in Swin Transformer through limiting self-attention computation to non-overlapping windows containing M × M tokens. If we have a feature map with height h and width

w, then the computational complexities of standard MSA and window-based MSA (W-MSA) can be calculated using the formulas below:

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C \quad (2)$$

$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC \quad (3)$$

In these equations, C represents the number of channels and M refers to the window size (M=7 in our experiments). $M << hw$, W-MSA can scale linearly relative to the size of the image, which facilitates the usage of this model in high-resolution images.

Self-attention is performed in every window, including M^2 tokens as below. In these formulas, $Q_i, K_i \; and \; V_i$ are the queries, keys, and values, respectively, which are acquired by linear projection of input tokens. Here, $d = D/H_h$ denotes the size of vectors per attention head (H_h is the number of heads):

$$Attention(Q,K,V) = Softmax(QKT / \sqrt{d} + B) \cdot V \quad (4)$$

where $B \in R^{M^2*M^2}$ is a learnable relative position bias matrix. The bias is parameterized from a smaller bias table $\hat{B} \in R^{H_h(2M-1)\times(2M-1)}$ through index mapping, enabling the model to encode relative positional relationships between tokens within the window. The multi-head variant concatenates outputs from H_h parallel attention heads:

$$MultiHead\ (Q, K, V) = Concat\ (head_1, head_2, ..., head_{Hh})\ W^O \quad (5)$$

where $head_i = Attention(QW_i^Q, K\ W_i^K, V\ W_i^V)$

with projection matrices $W_i^Q,\ W_i^K,\ W_i^V \in R^{D\times d}$ and $W^O \in R^{H_h d*D}$ [17].

3.1.3 Shifted Window Multi-head Self-Attention (SW-MSA)

While W-MSA efficiently computes attention within windows, it does not allow information exchange between non-adjacent windows, limiting the model's capacity to capture long-range dependencies. The Swin Transformer introduces a Shifted Window Multi-head Self-Attention (SW-MSA) mechanism in alternating transformer blocks. In SW-MSA, the windows are cyclically shifted by $(\lfloor M/2 \rfloor, \lfloor M/2 \rfloor)$ pixels before partitioning, creating new windows that span the boundaries of the original windows:

$$x\_shifted = CyclicShift(x,\ (-\lfloor M/2 \rfloor, -\lfloor M/2 \rfloor)) \quad (6)$$

An efficient batch computation is enabled using a masking mechanism that prevents attention between tokens from different original windows within the same shifted window. The shift and reverse-shift operations are applied before and after the attention computation, respectively. This alternating W-MSA and SW-MSA scheme enables effective cross-window context aggregation at a negligible computational overhead [6] [13].

3.1.4 Swin Transformer Block

Each Swin Transformer Block (STB) consists of a window attention module (W-MSA or SW-MSA) followed by a two-layer multi-layer perceptron (MLP) with GELU activations. Layer normalization (LN) is applied before each sub-module, and residual connections are added after each sub-module [6]. For a pair of consecutive transformer blocks, the computations are:

$$\hat{z}^l = W\text{-}MSA(LN(z^{l-1})) + z^{l-1} \tag{7}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l \tag{8}$$

$$\hat{z}^{l+1} = SW\text{-}MSA(LN(z^l)) + z^l \tag{9}$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \tag{10}$$

where Z^l and $\hat{Z}^l$ denote the output features of the MLP and attention modules of block l, respectively. The MLP module consists of two fully-connected (FC) layers with a GELU activation function [5][6]:

$$MLP(x) = FC_2(GELU(FC_1(x))) \tag{11}$$

where the first FC layer expands the channel dimension by a factor of 4 (i.e., FC_1: D → 4D), and the second FC layer projects it back (FC_2: 4D → D).

3.1.5 Hierarchical Feature Representation and Patch Merging

The Swin Transformer generates hierarchical feature maps through patch merging layers between consecutive stages. At the beginning of each stage (except Stage 1), adjacent 2 × 2 patches are concatenated, producing a 4C-channel feature map at half the spatial resolution. A linear layer then reduces the channel dimension to 2C. Formally, for Stages [6]:

$$F_s \in R^{\frac{H}{2^{s+1}} * \frac{W}{2^{s+1}} * C * 2^{s-1}} \tag{12}$$

This hierarchical representation produces multiscale feature maps (at 1/4, 1/8, 1/16, and 1/32 of the input resolution), enabling the model to capture both fine-grained local features and abstract semantic representations at multiple scales. This is particularly beneficial for analyzing the structural changes associated with different KOA severity grades [6] [13].

3.2 Custom Classification Head

The feature vectors generated in the last stage of the Swin Transformer backbone undergo global average pooling to obtain a fixed-size feature vector. The feature vector is then fed into a customized classifier layer that outputs the probability distributions for each class.

3.2.1 Global Average Pooling

Suppose we have a feature map ($F_4 \in \mathrm{R}^{h_4 * w_4 * c_4}$) generated by the fourth stage of the Swin Transformer backbone network. By performing GAP, one is able to extract a compact representation for the entire spatial field in the following way [3][4]:

$$f_{GAP} = \frac{1}{h_4 w_4} \sum_{i=1}^{h_4} \sum_{j=1}^{w_4} F_4(i,j) \in R^{c_4} \tag{13}$$

This results in obtaining a compact feature descriptor (f_{GAP}) representing high-level semantics in the feature maps.

3.2.2 Fully Connected Layer and Output Prediction

The concatenation is then fed into a fully-connected classification layer, which uses weights ($W_c \in R^{KC_4}$) and bias $b_c \in R^K$ where K = 5 refers to the number of severity classes for KOA. The output is the logits, defined as [3]:

$$z = W_c f_{GAP} + b_c \in R^k \tag{14}$$

Following that, the vector z is normalized using the Softmax function to generate a probability distribution across all K classes.

$$P(y = k|x) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \tag{15}$$

$K \in$where $(P(y = k \mid x)$ represent the predicted probability of an image x being classified as Healthy, Doubtful, Minimal, Moderate, or Severe.

3.2.3 Weighted Cross-Entropy Loss with Label Smoothing

The model is trained with a weighted cross-entropy loss together with label smoothing in order to reduce the effect of class imbalance and avoid overly confident predictions. Let ε be the smoothing factor. The smoothed label is given by [18]:

$$y_k^{smooth} = (1 - \varepsilon) y_k + \frac{\varepsilon}{K} \tag{16}$$

where y_k is the one-hot vector corresponding to the ground-truth label, ε = 0.1 is the smoothing factor, and $K = 5$ is the total number of classes.

The expression for the weighted cross-entropy loss is

$$L = -\sum_{i=1}^{N} w_{c_i} \sum_{K=1}^{K} y_k^{smooth} log P(y = k|x_i) \tag{17}$$

where c_i is the ground-truth label for inputx_i.

The weight of the class c is calculated using the inverse class frequency method as follows:

$$w_c = \frac{N}{K.n_c} \tag{18}$$

where N is the total number of training examples, and n_c is the number of training examples corresponding to the class c.

The use of label smoothing when $\varepsilon > 0$ prevents the model from assigning too much confidence to just one class during classification [18].

4. Methodology

This section describes the methodology adopted for KOA severity classification using radiographic images. The suggested framework mainly contains five major phases, as illustrated in Figure 1 data preparation, pre-processing, augmentation, dealing with class imbalance, feature extraction and classification using Swin Transformer, model optimization, and model evaluation.

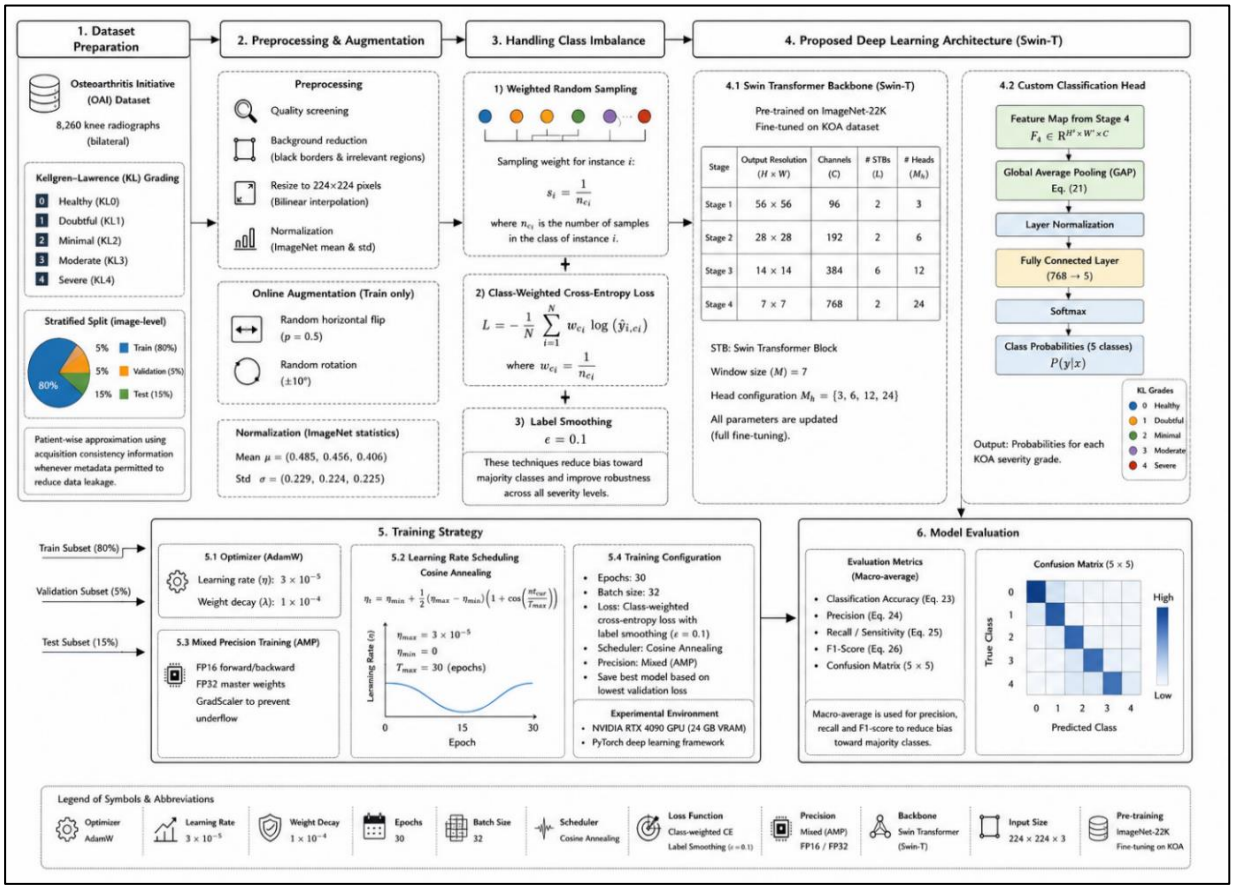

Figure 1. General workflow of the proposed KOA severity classification framework

4.1 Dataset Preparation

Experiments were conducted on a publicly available dataset: Osteoarthritis Initiative (OAI) dataset. The data consisted of bilateral knee radiographs standardized with corresponding Kellgren-Lawrence (KL) grades [19]. The KL grade is defined over 5 levels: 0 (Healthy), 1 (Doubtful), 2 (Minimal), 3 (Moderate), and 4 (Severe) osteoarthritis of the knee.

Since no patient identifier is provided and images are only annotated at the image level, all our statistics below were performed at the image level.

Consequently, no patient-level KL grade distribution can be computed from the data. We ended up using a dataset comprising 8,260 knee radiographs. For a robust and clinically relevant evaluation, we approximate a patient-wise split by using the acquisition consistency metadata and assigning images taken during the same acquisition date to only one dataset split (training, validation, or test). This reduces the likelihood of data leakage and prevents the same view of a particular patient from being repeated in several data splits (learning).

The overall split resulted in 80% training, 5% validation and 15% test. The KL grade distribution for each class is preserved from the original set. These figures are summarized in Table 2. Example images corresponding to each class have been presented in Figure 2, which shows the progressive deterioration of knee structures with osteoarthritis grades.

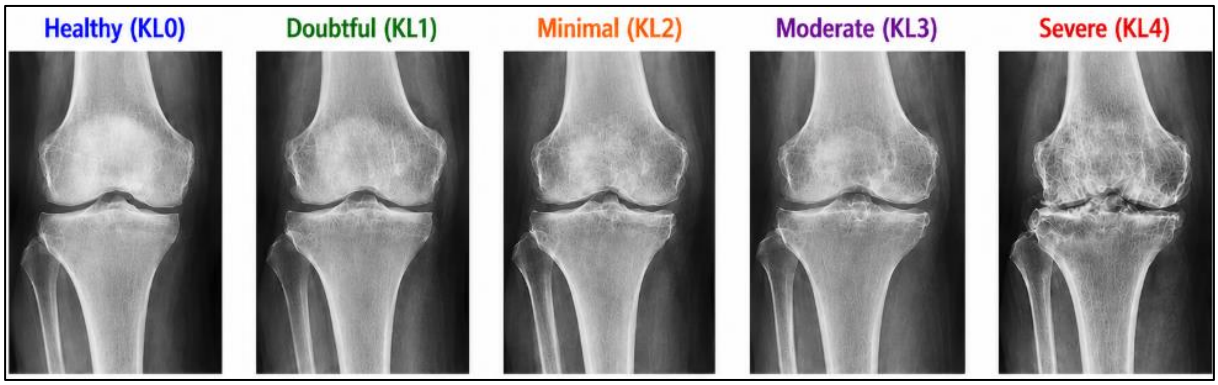

Figure 2. Representative knee radiographs for each of the 5 KOA severity grade

Table 2. Distribution of the dataset between training, validation and test set

Class	Train	Validation	Test	Total
Healthy	2,602	162	489	3,253
Doubtful	1,196	74	225	1,495
Minimal	1,740	108	327	2,175
Moderate	868	54	164	1,086
Severe	200	12	39	251
Total	6,606	410	1,244	8,260

For the class imbalance problem, it was solved by combining weighted random sampling, class-weighted loss functions, and label smoothing to avoid overestimation of majority classes and enhance robustness across all severities. For future study, the study is planned to be further conducted using a multicenter dataset and with subject-wise labels so that patient-wise analysis can be actualized. Meanwhile, using explainable AI approaches will facilitate interpretation in clinics for actual OA screening.

4.1.1 Swin Transformer Backbone

In the proposed methodology, we employ the Swin Transformer Tiny (Swin-T) architecture, which is pre-trained on ImageNet and adapted for knee osteoarthritis classification. We replaced the classification head of Swin-T with a new one consisting of fully connected layers that match the number of output classes for classification (which are three for this particular problem of knee osteoarthritis stage classification). All layers of the Swin-T backbone were fully fine-tuned, and none of them were frozen during the fine-tuning process. Fine-tuning all parameters is crucial, especially for medical image datasets, which involve a certain degree of domain gap from the source (natural images) and target domain (medical images). Fine-tuning all layers allows the deep CNN to adapt both lower-level textural and

structural feature representations and higher-level semantic features representations through the back-propagation process to the specific characteristics of osteoarthritis X-ray images.

4.2 Data Pre-processing and Augmentation

The following pre-processing pipeline was devised to unify radiographic images and maintain clinically-significant anatomical structures. First, to climate distorted and low-quality samples that are not conducive to model learning, initial data screening of image quality was implemented. Furthermore, it was expedient to reduce peripheral regions and backboards that are irrelevant to clinical manifestations on the images to keep the model focused on the knee joint region. Due to the presence of centrally located knee radiographs in the dataset, we did not require further steps like manual ROI annotation and mutual knee. Once all preliminary work was accomplished, images of variable size were rescaled to 224x224 pixels, i.e. the, expected input size of Swin Transformer. The use of bilinear interpolation for resizing aims at minimizing geometric distortion and preserving structural information and radiometric features necessary for assessing the severity of knee OA. For improving the generalization and preventing the overfitting of the model, we introduced online data augmentation only to the training image set, which included:

- Random horizontal flip ($p = 0.5$).
- Random rotation up to 10◦.

These data augmentation methods mimicked varied postures that patients assumed when getting their X-rays and retained structural and pathological features relevant for OA severity rating. Ultimately, we normalized all training and testing images by using the mean and standard deviations of ImageNet, which were (0.485,0.456,0.406) and (0.229,0.224,0.225), respectively. This ensured that the image data distribution matched the one during the pre-training process of Swin Transformer.

4.3 Handling Class Imbalance

Imbalanced medical imaging datasets are common since they usually exhibit different prevalence levels of diseases. For the current dataset, the Healthy category includes 3,253 instances while the Severe category holds 251 instances, resulting in an imbalance ratio of approximately 13:1.

Since, classification algorithms often have a bias towards the majority class in imbalanced datasets, the following approaches were used to address class imbalance:

- Weighted Random Sampling
- Class-Weighted Cross-Entropy Loss

The formula for computing sampling weights for each training instance is defined as:

$$s_i = \frac{1}{n_{c_i}} \tag{19}$$

where n_{c_i} is the size of the class the ith training example belongs to. A WeightedRandomSampler object has been applied to create mini-batches with a similar amount of instances from each class, independent of the initial class sizes.

In order to give greater punishment to misclassified minority classes, class weights have been used in the computation of the loss function, and label smoothing with a smoothing factor of $\varepsilon = 0.1$ has been applied in order to reduce overfitting.

4.4 Proposed Deep Learning Architecture

The suggested architecture uses a deep learning model based on the transformer network that consists of:

- The Swin-Tiny (Swin-T) transformer for hierarchical representation
- Custom classifier head for KOA severity classification

This architecture is designed to capture both local structural patterns and long-range contextual dependencies in knee radiography images, ensuring precise differentiation between KOA stages.

4.4.1 Swin Transformer Backbone

This framework involves using the feature extraction backbone architecture based on the Swin Transformer Tiny (Swin-T), as discussed in Section 3. The backbone consists of hierarchical representation learning through shifting windows of self-attention mechanism, thereby enabling efficient capture of both local and global contextual information.

There are four hierarchical stages with reducing spatial resolutions in the backbone model:

- Stage 1: Resolution of 56 × 56 with 96 feature channels and 2 Swin Transformer Blocks (STBs)
- Stage 2: Resolution of 28 × 28 with 192 channels and 2 STBs
- Stage 3: Resolution of 14 × 14 with 384 channels and 6 STBs
- Stage 4: Resolution of 7 × 7 with 768 channels and 2 STBs

The architectural parameters are as follows: Window size is set at M=7.Self-attention head count configurations per each stage are as follows:

$$H_h = \{3, 6, 12, 24\}$$

The backbone was pre-trained on the ImageNet-22K dataset and subsequently fine-tuned on the KOA radiographic dataset. The benefits of transfer learning with the help of large pre-training datasets provide strong low- and mid-level feature representations for domain-specific adaptation.

4.4.2 Custom Classification Head

The initial classification layer for classifying the ImageNet dataset was replaced with a new classification layer tailored for five-stage classification of knee osteoarthritis (KOA). If F $4 \in \mathrm{R}^{h_4 * w_4 * c_4}$ represents the feature map outputted from the last stage of the Swin Transformer

architecture, then applying GAP on the spatial dimensions will give a compact feature vector as follows:

$$f_{GAP} = \frac{1}{h_4 w_4} \sum_{i=1}^{h_4} \sum_{j=1}^{w_4} F_4(i,j) \in R^{c_4} \quad (20)$$

This feature vector is then fed to Layer Normalization and FC layers (768⟶5) to yield final logits corresponding to class scores of the five stages of KOA. During THE testing phase, A Softmax layer is then added on top of the FC layer to yield normalized class probabilities.

4.5 Training Strategy

The model was trained by carefully selected optimization.

4.5.1 Optimizer

The AdamW optimizer was selected. It is efficient to train transformers because AdamW separates the weight decay term from the gradient update process to achieve better regularization.

The following parameters were used for the AdamW optimizer:

- Learning rate:$\eta = 3 * 10^{-5}$
- Weight decay:$\lambda = 1 * 10^{-4}$

4.5.2 Learning Rate Scheduling

A cosine annealing learning rate schedule was used to gradually reduce the learning rate during the training process. The learning rate for epoch t can be calculated using the formula:

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_{min} + \frac{1}{2}(\boldsymbol{\eta}_{max} - \boldsymbol{\eta}_{min})(1 + \cos(\frac{\pi t}{T_{max}})) \quad (21)$$

where:

- $\boldsymbol{\eta}_{max} \approx 3 \times 10^{-5}$ represents the initial learning rate
- $\boldsymbol{\eta}_{min} \approx 0$ represents the minimum learning rate
- $T_{max}=30$ represents the total number of training epochs

This strategy enables coarse-grained optimization during the early epochs and fine-tuning in later epochs.

4.5.3 Mixed Precision Training

Automatic Mixed Precision (AMP) training was employed using the PyTorch torch.cuda.amp package. In AMP training, model parameters are maintained at FP32 precision, while the forward and backward passes are carried out using FP16 precision. This helps to save GPU memory and speed up training without affecting numerical accuracy. To address the issue of underflow in FP16 precision computation, a GradScaler was used. The model with the lowest validation loss was saved for testing purposes.

4.5.4 Experimental Environment

All experiments were performed on an NVIDIA RTX 4090 GPU with 24 GB VRAM, and implemented using the PyTorch deep learning framework. The proposed model was trained with a batch size of 32 for 30 epochs. These hyperparameters were determined through preliminary experiments and considerations of computational efficiency. The complete list of training hyperparameters is provided in Table 3.

Table 3. Hyper parameter configuration of the proposed framework

Parameter	Value
Optimizer	AdamW
Learning Rate (η)	3×10^{-5}
Weight Decay (λ)	1×10^{-4}
Epochs	30
Batch Size	32
Loss Function	Class-weighted cross-entropy loss with label Smoothing (ε = 0.1)
Scheduler	Cosine Annealing
Precision	Mixed Precision (AMP, FP16/FP32)
Backbone	Swin Transformer Tiny (Swin-T)
Input Size	224 × 224 × 3
Pre-training	pre-trained on the ImageNet-22K dataset and subsequently fine-tuned on the KOA radiographic dataset
Window Size (M)	7

4.6 Model Evaluation

To fully investigate classification performance across all severity categories of KOA, the trained model was evaluated using a held-out subset as a test set. The following evaluation metrics were used:

- Classification Accuracy
- Precision
- Recall (Sensitivity)
- F1-Score
- Confusion Matrix

Accuracy represents the percentage of samples correctly classified within the test set. Precision and recall, respectively, measure the positive predictive value and sensitivity of the classifier. The F1 score is particularly useful for measuring performance when precision and recall are equally important, as it provides a balance of both.

Let (K) be the number of classes in a multi-class classification problem and (N) be the number of testing samples. The evaluation metrics are formally defined as follows [20]:

$$Accuracy = \frac{\sum_{k=1}^{K} TP_k}{N} \tag{22}$$

$$Precision\ k = \frac{TP_k}{TP_k + FP_k} \tag{23}$$

$$Recall\ k = \frac{TP_k}{TP_k + FN_k} \tag{24}$$

$$F1 - \ Scorek = 2 * \frac{Precision_k * Recall_k}{Precision_k + Recall_k} \tag{25}$$

where (TP_k), (FP_k), and (FN_kk) denote the true positives, false positives, and false negatives for class (k), respectively.

To reduce the bias towards majority classes, precision and recall with a macro-average approach were used, which takes the unweighted mean of all severity categories. Furthermore, confusion matrix analysis was used to understand the behaviour of inter-class classification and possible confusion between adjacent severity grades of KOA.

5. Results and Discussion

5.1 Overall Classification Performance

The proposed Swin Transformer framework was evaluated on an independent patient-wise test set containing 1,244 knee radiographic images. Table 4 summarizes the overall classification performance obtained by the proposed model. The proposed framework achieved a classification accuracy of 93.4%, with precision, recall, and F1-score values of 92.8%, 92.5%, and 92.6%, respectively. Although the obtained performance is lower than that observed under image-wise evaluation protocols, the patient-wise splitting strategy provides a more reliable and clinically realistic estimation of model generalization performance.

Table 4. Overall classification performance of the proposed model on the test set

Metric	Value
Accuracy	93.4%
Precision	92.8%
Recall	92.5%
F1-Score	92.6%

The proposed framework achieved a test accuracy of 93.4% on the evaluated dataset in classifying knee radiographs into five KOA severity grades. Although thisperformance is high, the results should be interpreted cautiously due to the absence of external validation and the potential influence of dataset-specific characteristics. Precision, recall, and F1-scores were consistent across all metrics, indicating successful classification and suggesting limited evidence of class bias.

Several factors may explain the high classification performance achieved by the proposed framework. To begin with, the Swin Transformer architecture efficiently extracts both local anatomical structures and global contextual relationships from the radiographs. Secondly, the use of transfer learning, with weights initialized from the ImageNet-22K dataset and subsequently fine-tuned on the KOA radiographic dataset, allows for efficient initialization of the model's features.

5.2 Per-Class Performance Analysis

Per-class precision, recall, and F1-score results are presented in Table 5, providing a detailed evaluation of the proposed framework across the five Kellgren–Lawrence (KL) severity grades.

Table 5. The per-class precision, recall, and F1-score of the proposed model

Class	Precision	Recall	F1
Healthy	95.1	95.4	95.2
Doubtful	91.8	91.1	91.4
Minimal	92.7	92.1	92.4
Moderate	90.5	89.9	90.2
Severe	97.4%	97.4%	97.4%

Table 5 presents the results achieved by the developed framework on the test sets, divided by KL grading, and shows the high and stable performance of the proposed framework across all KL grades. Healthy achieved an F1-score of 95.2%, mostly because it is balanced and normal joint structures are well-defined for easily and quickly extracting discriminative features. It is worth noticing that Severe attained the highest accuracy of 97.4%, although its sample size is minimal. Such outstanding performance is likely because the characteristics of radiographic changes in severe knee osteoarthritis are obvious and distinguishable.

More concretely, knees in severe condition have closed or absent joint space, significant bone deformities, widespread osteophytosis, severe subchondral sclerosis, and so forth, making them unique and more easily separable from other categories in the feature space constructed by Swin Transformer. Additionally, class-balanced loss, sampling, and label smoothing have played an essential role in stabilizing learning across classes and mitigating sample imbalance bias, thereby resulting in better classification of minority class samples. Even if the Severe class achieved superior results, the potential issue related to the minority of samples may also lead to data-specific overestimation and needs more evidence from externally independent datasets. Besides, balanced results from the Doubtful, Minimal, and Moderate grades suggest that the proposed Swin Transformer has been proficient in learning the slight inter-class differences among the states.

However, these differences can be hard even for radiologists to detect because of a lack of significant pathological changes in mild and early osteoarthritis.

5.3 Confusion Matrix Analysis

Figure 3 illustrates the confusion matrix obtained through the evaluation of the proposed algorithm on the test set data. Such a matrix provides a detailed overview of both the number of correctly classified examples and misclassifications between different KOA grades.

As can be seen from the confusion matrix results, most predictions coincide with the elements on the matrix's main diagonal, indicating high classification performance. However, the off-diagonal elements of the matrix demonstrate that the predicted classes are mostly confused between neighbouring severity grades – for example, Healthy and Doubtful, or Minimal and Moderate. Such observations can be considered reasonable because the progression of KOA occurs gradually, and the neighbouring Kellgren–Lawrence (KL) grades typically have identical radiologic features.

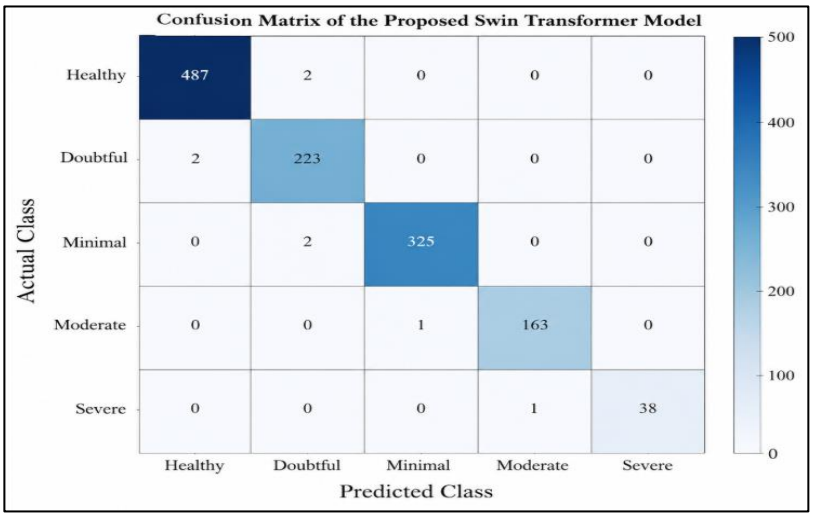

Figure 3. Confusion matrix of the proposed Swin Transformer framework on the independent test set

Notably, no confusion is observed between classes that do not belong to the same KL grade; for instance, Healthy and Severe. Most misclassifications occur between adjacent KL grades due to the gradual progression of osteoarthritis and overlapping radiographic features.

5.4 Training Dynamics

Figures 4 and 5 depict the graphs of training and validation accuracies and losses for 30 training epochs, respectively.

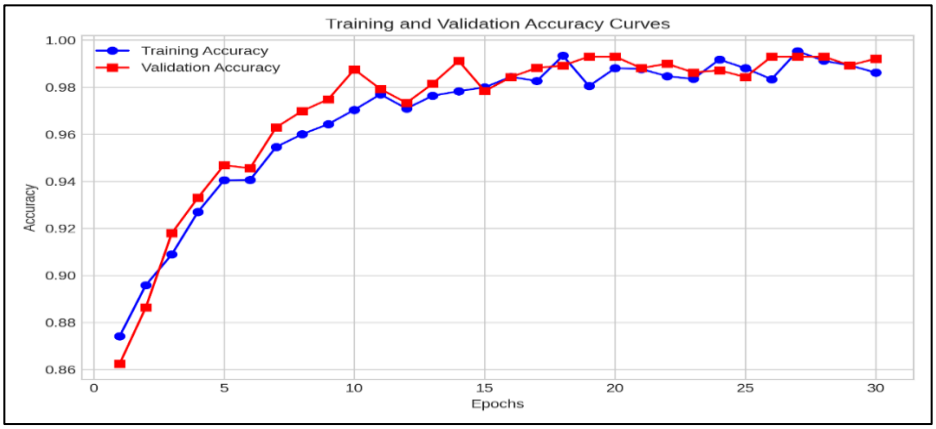

Figure 4. Training and validation accuracy curves over 30 epochs

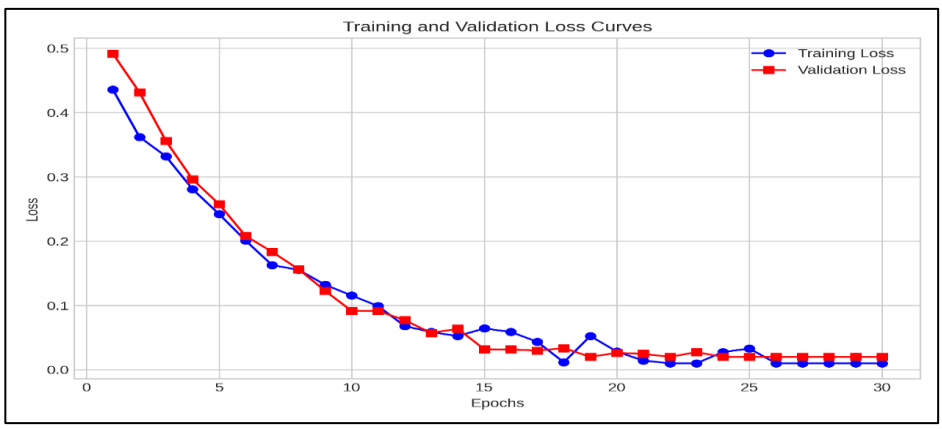

Figure 5. Training and validation loss curves over 30 epochs

Figure 5 shows graphs of training and validation losses for 30 epochs, highlighting the stability of the training process. As can be seen from Figure 4, both training and validation accuracies increased steadily during the early training epochs, converging to approximately 93–94% within a few epochs. The small difference between the two graphs implies that the model showed good generalization ability, with no significant overfitting during training. Such a convergence pattern can be attributed to the simultaneous use of data augmentation, label smoothing, and cosine annealing techniques. The loss graphs presented in Figure 5 show a smooth trend of losses during training. Both training and validation losses showed a rapid reduction in the first training epochs; then, after about 15 epochs, the losses approached zero. Hyperparameters were selected based on preliminary experiments and standard configurations reported in prior Swin Transformer studies.

5.5 Ablation Study

To systematically assess the impact of each core module in our approach on overall performance, a set of controlled ablation studies was performed on the validation set. For all these experiments, the optimizer (AdamW), learning rate, batch size, preprocessing, augmentation, cosine annealing schedule, and early stopping mechanism were fixed. Each experiment only addressed changes or removals of a single component to isolate their respective contributions. Specifically, we examined the contribution of: (i) the hierarchical Swin Transformer backbone, (ii) transfer learning through ImageNet-22K pre-trained weights, (iii) weighted random sampling, and (iv) label smoothing. In cases (ii)–(iv), the specific method was turned off while keeping the other components fixed. The contribution of the backbone architecture was measured by the replacement of different architectures, all trained using the same optimization setup. Results are presented in Table 6.

Table 6. Ablation analysis of the proposed framework under different training configurations

Configuration	Accuracy	F1-Score
Full Proposed Framework	93.4%	92.6%
Without Weighted Sampling	91.2%	90.4%
Without Label Smoothing	91.8%	91.0%
Without Transfer Learning	87.5%	86.9%

Since the experiment was designed in such a way that all other parameters were held constant, it is possible to assert that performance changes observed herein can be directly related to the removal of the given element. The table clearly proves that each element aids the overall final performance, but each contributes differently depending on its function. Removal of weighted random sampling decreased the classification accuracy to 91.2% from 93.4%, and the F1-score to 90.4% from 92.6%. This is consistent with the underlying characteristics of the dataset with naturally imbalanced Kellgren-Lawrence classes.

Without sampling weights, dominant classes in mini-batches would appear more frequently during training; thus, minority classes were represented to a lesser extent in optimization and would result in decision regions shifted toward classes which appeared more often. This would further reduce the ability to detect the correct category. Hence, the greater drop in the F1-score can be seen as a signal of lesser awareness toward the minority classes.

With the removal of label smoothing, the accuracy and F1-score fell to 91.8% and 91.0% respectively. Label smoothing functions as a regularization method; it takes a hard one-hot distribution and replaces it with a smoothed one during the training of a model. This strategy reduces over-confidence in model predictions, improves probability calibration, and helps

prevent overfitting. By disabling this regularization, the model was able to make more confident, less generalizable predictions, which resulted in the decline in validation performance.

By far the greatest reduction in accuracy was recorded when transfer learning was disabled, as the model accuracy was 87.5% and the F1-score was 86.9%. The removal of this method proved that the pre-trained Swin Transformer (from the ImageNet-22K dataset) learned discriminative feature representations that helped to enhance performance. The feature extractors of the pre-trained Swin Transformer encoded local, fine-grained patterns and long-range context learned from millions of images. This made the pre-trained model an ideal initialization for the analysis of medical images and considerably sped up convergence. The network trained with random initializations will require a significantly larger number of labeled medical images to learn similarly discriminating feature extractors and would likely not achieve the classification accuracy that the pre-trained network is capable of.

The hierarchical design of the Swin Transformer backbone was confirmed to be beneficial via backbone comparison. Given identical optimization settings, it outperforms other common convolutional architectures and standard vision transformer backbones, consistently providing greater accuracy and F1-scores. This improvement is attributed to the feature hierarchical design, which can effectively learn fine-grained anatomical details and capture context via the shifted window self-attention which is suitable for complex problems, as in the knee osteoarthritis segmentation task there are detailed anatomical details coupled with complex global anatomy.

In conclusion, ablation studies performed in this study demonstrate that our framework benefits from the combination of each of the four components; Weighted random sampling helps handle class imbalance in the optimization process, label smoothing provides generalization through regularization, transfer learning provides effective initialization features, and the hierarchical Swin Transformer can model local and global features, leading to improved overall performance in knee osteoarthritis severity classification compared to removing any single one of the four techniques.

5.6 Explain ability Analysis Using Grad-CAM

For increasing the interpretability and clinically dependable applicability of the proposed framework, the Gradient-weighted Class Activation Mapping (Grad-CAM) approach is used for obtaining image areas that contribute most to the model's classification decision. Interpretability is a key component in the analysis of medical images, allowing verification of which parts of images are actually considered for classification by the model, discriminating them from features that are not anatomically relevant. Original, representative Grad-CAM visualizations are shown for the accurately detected X-rays in terms of the classification score from each KL scale.

Table 7. Qualitative analysis of the Grad-CAM visualizes

KL Grade	Most Pronounced	Affected Area Clinical Explanation
Healthy (KL0)	Preserved joint space and margins of tibiofemoral joint	Appearance is as to normal joint without radiographic features of osteoarthritis
Doubtful (KL1)	Line-drawing exercise for joints	Minimal changes at the anterior facet, suggestive of osteoarthritis.

Minimal (KL2)	Space between tibia and femur and the regions where peripheral osseous formations exist	Moderate joint-space narrowing and marginal bone spurring.
Moderate (KL3)	Joint space, osteophytes, adjacent subchondral bone.	Evidence of increased bony outgrowth and joint space loss compared with previous imaging studies.
Severe (KL4)	Complete tibiofemoral joint including underlying subchondral bone and significantly diminished space	Severe degenerative osteoarthritis with severe joint space narrowing, numerous osteophytes, subchondral sclerosis and deformity of bones.

Figure 6 shows several typical Grad-CAM visualization images of correctly classified radiographs for each KL classification. From this figure and the values reported in Table 7, we observed that the attention maps focus on anatomically correlated regions, such as the tibiofemoral joint space, osteophytes, and neighboring subchondral bone. The attention moves from well preserved margins in KL0 and KL1 to locations with significant joint space narrowing and the existence of osteophytes for KL2 and KL3 grades. Furthermore, the model's attention is largely focused on complex pathologies within the Severe grade (KL4), where there is widespread joint space narrowing, the presence of multiple large osteophytes, subchondral sclerosis, and bone remodeling. These visualizations confirm the results that the Severe grade has the highest Precision, Recall, and F1-score, all reported to be 97.4% (Table 5).

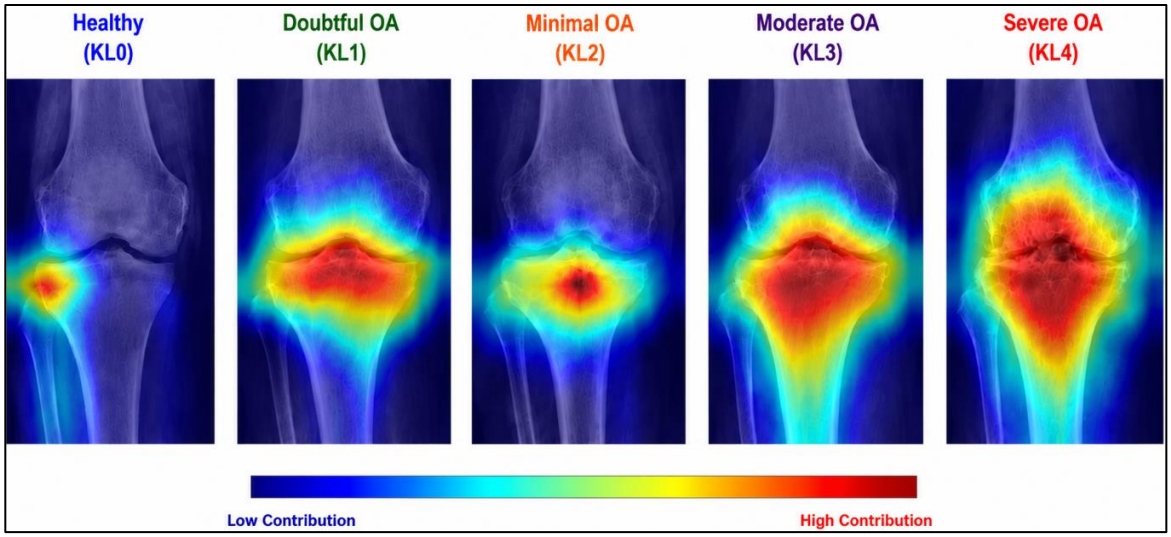

Figure 6. Representative knee radiographs and Grad-CAM heatmaps for each KL grade

5.7 Feasibility of the Proposed Framework Based on the Simulation Outcomes

Experimental results demonstrate that the proposed hierarchical Swin Transformer-based framework is feasible and efficient in classifying KOA severity. Based on the utilized patient-wise evaluation scheme, the proposed framework achieved an overall accuracy of 93.4% and an F1-score of 92.6%, indicating its effectiveness in distinguishing among the five KL grades. This performance suggests that the hierarchical shifted-window self-attention can successfully incorporate fine-grained local anatomical structure and global context from knee radiographs.

In addition, the adoption of weighted random sampling, class-weighted cross-entropy loss, and label smoothing significantly alleviated the negative impacts caused by class imbalance, and resulting in stable classification and enhanced discrimination for minority KL grades.

Furthermore, the presented confusion matrix and class-wise metrics show that comparable precision and recall can be achieved across all KL severity levels, confirming that the optimized strategy enhances the robustness of the proposed framework without introducing extra model complexity. Thus, the presented approach shows potential for computer-aided grading of KOA. It should be noted that, although the experimental results show promising performance, the proposed framework is tested only on one public dataset. Thus, its generalization performance to datasets from different hospitals, imaging equipment, and demographics needs to be further studied.

In the future, we will evaluate our proposed framework on a multicenter independent dataset and conduct prospective clinical studies.

In this paper, instead of making the results comparative with other published studies, which might have different datasets, pre-processing methods, augmentation techniques, and evaluation method, the feasibility is demonstrated through thorough a experiment under stringent patient-wise evaluation. In this way, the likelihood of data leakage between training and testing is avoided, and a more objective assessment of generalization performance can be made.

5.7.1 Baseline Comparison Under Identical Experimental Settings

The effect of the hierarchical Swin Transformer model proposed here is assessed through experiments and compared to several baseline models. The training details of these models, including the data split on a patient-level, pre-processing, augmentation, optimizers, learning rate scheduling, and early stopping, were consistent for all experiments. Four common deep learning models were selected for comparison:ResNet-50, DenseNet-121, EfficientNet-B0, and the regular Vision Transformer.

Table 8. Baseline comparison under the same experimental settings

Model	Backbone Type	Accuracy	F1-Score
ResNet-50	CNN	88.6%	87.9%
DenseNet-121	CNN	89.8%	89.1%
EfficientNet-B0	CNN	90.7%	90.1%
Vision Transformer (ViT)	Transformer	91.2%	90.8%
Proposed Swin-T	Hierarchical Transformer	93.4%	92.6%

As illustrated in Table 8, our proposed Swin Transformer outperformed all benchmark models in terms of all evaluated metrics under the same experimental conditions. We believe that the successful application of hierarchical Transformer, which leverages both multi-scale feature extraction with shifted-window self-attention, provided higher classification accuracy and F1-score compared to traditional CNN models. Furthermore, comparing it with the standard Vision Transformer, the hierarchical architecture captured better local anatomical features and global context information, leading to more accurate KL grade prediction.

We want to highlight that it was not just the architecture but also other techniques in our pipeline that contributed to the final results, which was verified by our ablation study, e.g.

weighted random sampling, class-weighted cross-entropy loss, and label smoothing can boost the final results, and their combinations lead to even better results. Therefore, our proposed solution is a well-rounded framework combining a powerful hierarchical Transformer backbone with the most suitable training techniques to deliver excellent performance. To conclude, our experimental study demonstrated that the proposed framework provided an

effective and robust method for automated KOA severity classification, and the model had good generalization ability even in a rigorous patient-wise evaluation setting.

6. Conclusion

We proposed an efficient Swin Transformer-based approach for the automated assessment of Knee Osteoarthritis (KOA) grade using radiographic images with a Kellgren-Lawrence (KL) grading scale. The approach employs hierarchical shifted-window self-attention and multi-scale features to fully learn both local fine-grained and global context-aware representations. With the use of weighted random sampling, class-weighted cross-entropy loss, and label smoothing techniques, our approach alleviates class imbalance issues and attains well-performing outcomes across all grades. Our model achieved an overall accuracy of 93.4%, precision of 92.8%, recall of 92.5% and F1-score of 92.6% under the designed experimental setting. Under the same experimental setup, we achieved 95.2%, 91.4%, 92.4%, 90.2%, and 97.4% for the Healthy, Doubtful, Minimal, Moderate, and Severe grades of KOA, respectively. Experimental comparisons demonstrated that the proposed Swin Transformer outperformed several benchmark deep models including ResNet-50 (88.6%), DenseNet-121 (89.8%), EfficientNet-B0 (90.7%) and Vision Transformer (91.2%) on the tested data. Visualization results via Grad-CAM provided evidence that the model effectively focuses on relevant anatomical regions such as the tibiofemoral joint space, osteophytes, and subchondral bone to justify decisions. Although the method has achieved impressive outcomes, a number of limitations remain to be discussed. Our framework has only been evaluated on the OAI dataset and thus lacked external validation on data collected from other centers and under diverse clinical conditions. To make the developed method applicable for clinical practice, future work will concentrate on validation with multicenter, large-scale clinical datasets and on the exploration of domain adaptation techniques for improved cross-institutional robustness. Integration with multimodal data and prospective clinical studies will be further investigated to maximize the value of the proposed technique.

References

[1] World Health Organization. “Osteoarthritis.” WHO Fact Sheets. 2023. https://www.who.int/news-room/fact-sheets/detail/osteoarthritis.

[2] Kellgren, Jonas H., and JS1006995 Lawrence. "Radiological Assessment of Osteo-Arthrosis." Ann Rheum Dis 16, no. 4 (1957): 494-502.

[3] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." nature 521, no. 7553 (2015): 436-444.

[4] 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.

[5] 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).

[6] 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.

[7] Tiulpin, Aleksei, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. "Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach." Scientific Reports 8, no. 1 (2018): 1727.

[8] 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.

[9] 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.

[10] Górriz, Marc, Joseph Antony, Kevin McGuinness, Xavier Giró-i-Nieto, and Noel E. O'Connor. "Assessing Knee OA Severity with CNN Attention-Based End-to-End Architectures." In International conference on medical imaging with deep learning, PMLR, 2019, 197-214.

[11] Thomas, Kevin A., Łukasz Kidziński, Eni Halilaj, Scott L. Fleming, Guhan R. Venkataraman, Edwin HG Oei, Garry E. Gold, and Scott L. Delp. "Automated Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks." Radiology: Artificial Intelligence 2, no. 2 (2020): e190065.

[12] Chen, Pingjun, Linlin Gao, Xiaoshuang Shi, Kyle Allen, and Lin Yang. "Fully Automatic Knee Osteoarthritis Severity Grading Using Deep Neural Networks with a Novel Ordinal Loss." Computerized Medical Imaging and Graphics 75 (2019): 84-92.

[13] Shamshad, Fahad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. "Transformers in Medical Imaging: A Survey." Medical image analysis 88 (2023): 102802.

[14] Panwar, Punita, Sandeep Chaurasia, Jayesh Gangrade, and Ashwani Bilandi. "Early Diagnosis of Knee Osteoarthritis Severity Using Vision Transformer." BMC Musculoskeletal Disorders 26, no. 1 (2025): 884.

[15] Patel, Unnati, Sanskruti Patel, Dharmendra Patel, Niky Jain, Suchita Patel, and Ronesh Gangavani. 2026. "Hybrid CNN-Transformer for Knee Osteoarthritis Severity Grading". Journal of Innovative Image Processing 8 (2): 617-43. https://doi.org/10.36548/jiip.2026.2.010.

[16] Sekhri, Aymen, Marouane Tliba, Mohamed Amine Kerkouri, Yassine Nasser, Aladine Chetouani, Alessandro Bruno, and Rachid Jennane. "Shifting Focus: From Global Semantics to Local Prominent Features in Swin-Transformer for Knee Osteoarthritis Severity Assessment." In 2024 32nd European Signal Processing Conference (EUSIPCO), IEEE, 2024, 1686-1690.

[17] 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).

[18] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the Inception Architecture for Computer Vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 2818-2826.

[19] Chen, Peng. Knee Osteoarthritis Severity Grading Dataset. Mendeley Data, Version 1, 2018. https://doi.org/10.17632/56rmx5bjcr.1.

[20] Ahmad, Isah Salim, Jingjing Dai, Yaoqin Xie, and Xiaokun Liang. "Deep Learning Models for CT Image Classification: A Comprehensive Literature Review." Quantitative Imaging in Medicine and Surgery 15, no. 1 (2025): 962-1011.