scientific reports

Check for updates

OPEN

Enhanced swin transformer with dual attention for knee osteoarthritis severity grading from X-ray images

K. Sudha[1]✉ & A. Rajiv Kannan[2]

Osteoarthritis of the knee (OA) is a common degenerative condition that affects quality of life and mobility, especially in older adults. For disease management and treatment planning to be successful, early and precise diagnosis is essential. In order to classify the severity of osteoarthritis (OA) from knee X-ray pictures, this study suggests a new hybrid deep learning framework called Swin-O-NETS. It combines a Fast Extreme Learning Network (FELN) with a Modified Swin Transformer with Multi-Headed Channel Self-Attention for feature extraction. For evaluation, 2,047 radiographs from five Kellgren–Lawrence (KL) severity classes were taken from the Osteoarthritis Initiative (OAI) dataset on Kaggle. Our model outperformed traditional CNN, ResNet, DenseNet, and ensemble methods, achieving state-of-the-art performance with 99.4% accuracy, 99.0% precision, 98.9% recall, 98.3% specificity, and 98.8% F1-score respectively. These findings show that the suggested approach, which has better robustness and less computational complexity, can consistently help with early OA grading. Future research will concentrate on integrating multimodal imaging data, producing lightweight versions for use in real-time healthcare systems, and testing the model on larger multi-center clinical datasets.

Keywords Knee osteoarthrits, Swin transformers, Fast extreme learning networks, Severity grades, Segmentation, Classification

One of the most common degenerative joint diseases in recent decades is knee osteoarthritis (OA)[1,2], which mostly affects the medial, lateral, and patellofemoral compartments of the knee and causes excruciating pain, stiffness, and impairment[3,4]. The World Health Organization estimates that 528 million people worldwide suffer from OA, with knee OA being the most prevalent type. According to epidemiological research, between 22 and 30 percent of those over 60 have knee OA, with women more likely than males to have the condition[5]. Nearly one in five persons over 50 in India have symptomatic knee OA, according to community-based surveys, making it one of the main causes of disability and a lower quality of life[6].

Osteophyte growth, joint space constriction, and articular cartilage degeneration are hallmarks of the disease's course, which ultimately results in chronic discomfort and functional restrictions. Age, obesity, hormonal imbalance, genetic susceptibility, prior joint trauma, and mechanical stress are all contributing factors to its complex etiology. The start and progression of OA are accelerated by the combination of these biological, environmental, and lifestyle-related risk factors. Patients may need a whole knee replacement in more advanced stages, which would be extremely expensive and time-consuming. Although behavioral interventions like exercise, muscle building, and weight control can reduce symptoms and delay the progression, they only provide short-term respite[7–9].

Since radiographic imaging is accessible, inexpensive, and safe, it continues to be the most popular diagnostic method for knee OA. The Kellgren–Lawrence (KL) grading system is the accepted diagnostic method. There are three main issues with radiographic interpretation, though: (i) subtle changes can cause early stages to be missed; (ii) diagnosis relies on subjective visual assessment, which can lead to inter-observer variability; and (iii)

[1]Department of Computer Science and Engineering (Cybersecurity), K S R College of Engineering, Tiruchengode, Namakkal District, Tamilnadu, India. [2]Department of Computer Science and Engineering, K S R College of Engineering, Tiruchengode, Namakkal District, Tamilnadu, India. ✉email: ksudhacse@outlook.com; srisudhan3@gmail.com

the lack of radiologists causes delays in clinical decision-making. The urgent need for automated, precise, and early diagnostic technologies that can help physicians take prompt action is highlighted by these limitations.

Recent developments in deep learning (DL) have demonstrated potential for use in OA categorization and other medical imaging applications. When it comes to knee radiograph analysis, Convolutional Neural Networks (CNNs) and their variations have demonstrated impressive performance[10–18]. However, their low capacity to capture long-range relationships, high computational cost, and high model complexity restrict their efficacy in early detection. As a result, the need for more effective and precise frameworks for OA severity classification is increasing.

To address these challenges, this study proposes a hybrid framework that integrates Modified Swin Transformers (MST) with Fast Extreme Learning Networks (FELN). While MST enhances feature extraction through multi-headed channel self-attention, FELN ensures efficient classification with reduced training overhead. The resulting Swin-O-NETS framework not only improves accuracy and reduces computational complexity but also demonstrates strong potential for early detection and grading of knee OA severity levels from radiographic images. By facilitating timely and precise diagnosis, the proposed approach aims to improve clinical decision-making and enhance patients' quality of life. The following is the paper's primary contribution:

1. To best of our knowledge, Transformer networks have not yet been used for the early classification of the OA based on the segmentation and feature extraction methods.
2. This proposed model introduces the Modified Swintransformer models and Fast Extreme Learning Networks to achieve the better classification performance of OA using X-ray images.
3. This model in cultivates the most prominent features by introducing the Multi-Headed Channel Attention layers before being sent to the FELN for classification.
4. The suggested models perform noticeably better for knee OA prediction than the most advanced DL techniques currently in use.
5. Furthermore, proposed network is also compared with the other variants of Swin transformer-based architectures in which the suggested methodology achieves the highest categorization efficiency.

The remainder of this paper is organized as follows. Section "Background views" provides the theoretical background of the Swin Transformer. Section "Proposed architecture" describes the materials, methodology, and techniques adopted in the proposed framework. Section "Experimental results and analysis" presents the experimental results along with their analysis and interpretation. Finally, Section "Conclusion and future discussion" concludes the paper and outlines directions for future research.

Related works

Numerous deep learning (DL) and artificial intelligence (AI)-based techniques have been put forth in recent years to enhance knee osteoarthritis (OA) diagnosis, prognosis, and monitoring. These studies have investigated a number of methods, such as transformer-based frameworks, CNN-based classification, joint space segmentation, and hybrid AI models, with encouraging outcomes across diverse imaging modalities like MRI and X-ray. Even with these developments, the majority of current approaches still have issues such high computational complexity, large memory needs, overfitting, lengthy training periods, or poor generalizability across various clinical datasets. The necessity for portable, precise, and effective open access grading systems that may be used in actual clinical settings is brought to light by a critical evaluation of these works. The following section reviews key contributions from recent literature, their strengths, and the limitations that motivate the development of the proposed Swin-O-NETS framework.

Shoaib et al. present a DL predictive methodology that predicts knee replacement (KR) with an AUC of 0.86 (p 0.05) using X-ray images with analytical and statistical information. Furthermore, the CNN Xception and Inception models are used to develop this predictive model. Accurate KR prediction models for clinical utility are identified by comparing the confusion matrix for both CNN models and the training and testing dataset's training and validation accuracy graphs. Its tremendous computational complexity, however, is this framework's fundamental flaw[19].

A DL model was suggested by Dalia et al.[20] to autonomously partition the knee region and utilize X-ray images to forecast when knee OA will start. Additionally, comparation analysis is given between segmenting knee joints utilisingan YOLOv5 and ensemble methodology. Several models of classification are evaluated for the KL grade categorization, such as ResNet and VGG16. Several experiments are carried out to clarify why the zone of interest segmentation stage is important for classifying KL grades. However, this strategy requires more training time.

The goal of Shen et al.[21] is to study a precise, memory-saving, and accurate CNN to accomplish the segmentation of the knee joints' medial and lateral joint spaces (LJS and MJS). RegNet and DeepLabV3P are used in this system to make predictions automatically. For the effective precision, a novel attentional mechanism known as the "Feature-Location" module was created. To create the final segmentations, the predictions were post-processed using alpha shape and area opening. The joint space segmentations yielded measurements of the volume, mean thickness, and standard deviation. These values are clinically significant for the diagnosis and monitoring of knee OA. However, it requires more memory for processing.

By using hybrid Artificial Intelligence (AI) methods, Janarthanan et al.[22] enables OA prediction to be automated without the need for professional help. In this study, a hybrid (SVM + NB) model as well as Linear Discriminant Analysis (LDA), Naive Bayes (NB), and Support Vector Machine (SVM) were used. AI algorithms need to be trained on tagged X-ray images to recognise OA. Therefore, Kaggle is employed to collect X-ray pictures of individuals with and without OA in several phases (Healthy, minimal, moderate, and severe). The AI model for training does not get the captured picture directly. In order to decrease the amount of memory

needed to train the AI model, feature extraction is done to the preprocessed picture. With regard to accuracy and memory use, this framework exhibits improved performance. However, the computational complexity is high.

Felfeliyan et al.[23] approach to segmenting unsupervised MRI data for knee osteoarthritis (OA) The CycleGAN model for MRI translation was released in 2021. According to this paradigm, our approach will help with the automatic unsupervised evaluation of knee MRI and free up specialists' the amount of time required for segmentation by hand otherwise by utilising frequently obtained MRI sequences. By automatically translating MRI scan to the comparable like it had been recorded using a different protocol or magnet, this technology provides dependable, automated assessment detached from hardware. As an example, routinely obtained, clinically collected knee MRI scans might be processed to produce high-contrast, high-resolution images suitable for automatic cartilage abnormality diagnosis. Although this framework's high-power consumption is a drawback, it is not the only one.

Viekash et al.[24] developed a DL-based control technique using CNNs as a means of actuating and controlling a Continuous Passive Motion (CPM) device. The patient's thigh muscles are equipped with EMG and IMU sensors, allowing the sensors to discriminate between three states of intent: advance, retrace, and relax. Regarding the technologies implementation, an affordable, green alpha prototyped CPM system is made. Three healthy volunteers were subjected to various experiments on which the data set was gathered. The effectiveness of the experimental results and precise CNN-based intuitive motion predictions demonstrate the viability of at-home rehabilitation equipment. However, the problems with over fitting remain with this framework.

Metrics of the Hip Knee Ankle Angle (HKAA) were predicted by Yan et al.[25] using radiography of the entire limb taken after Total Knee Arthroplasty (TKA) in patients. 2022 saw the construction and testing of deep learning models. After identifying 1899 radiography' Regions of Interest (RoI), the HKAA was computed and landmarks were identified using based on the collected ROI, regressed heatmaps were created. The attributes used to evaluate the system were the average and variance from each deviation between the annotations and the HKAA angle predictions. Among model projections and annotations, there was a postoperative HKAA variation ranging from 0.65° to 0.82° after surgery, with 95.0% of the difference being less than 1.5°. A fully automated method for measuring the HKAA on post-TKA full-limb radiographs of TKA patients. However, the significant computational complexity of this approach has been noted as a negative.

By using Stanford's MRNet Dataset, Azcona et al. offered a comparable assessment of old and new methodologies for identifying knee injuries in 2020. All methodologies are according to DL, and this framework compares the effectiveness of a deep residual network that has been created from scratch and transfer learning. With the use of more recent DL designs and methods for data augmentation, this system overall performed with an AUC of 93.4% on the validation data. The construction and training of models that analyse MRIs may be aided by the use of more adaptable architectures, which are currently being suggested. This approach discovered that the most important elements in selecting the optimum performance were transfer learning and a precisely calibrated data augmentation technique. The framework's slow processing speed, however, is its biggest flaw[26].

Supatman et al.[27] suggested using the Random Brightness Augmentation hyperparameter to classify with Deep CNN (DCNN), X-Ray Grade Kellgren-Lawrence (KL)-2 OA Initiative imaging datasets are processed. The study's findings showed that classification of X-ray narrowing pictures (KL-2) was capable of classifying images at any brightness, up to a value of 30, for the "Anterior View KOA" and "Posterior View KOA" categories, with training accuracy of 83.33% and validation accuracy of 54.69%. Yet this approach's significant computational cost is a key downside.

First, using a using radiograph and the Residual NN (ResNet), the knee joint was identified, according to Zhang et al. In order to automatically forecast the KL-grade, ResNet and Convolutional Block Attention Module (CBAM) were then combined. The suggested model much outperformed the published results, has a multi-class average accuracy of 74.81%, a quadratic Kappa score of 0.88, and a mean squared error of 0.36. In order to get knowledge about how the suggested model makes decisions, the attention maps were examined. nonetheless, issues with overfitting remain[28].

A 2023 study showed that, despite the need for extra imaging techniques and computer resources, combining multi-view radiographs with preexisting anatomical knowledge can greatly increase KL grading accuracy[29].

A hierarchical classification framework that included joint space and osteophyte segmentation was presented by Pan et al. in 2024[30]. This enhanced severity grading performance, but it also increased model complexity and required high-quality segmentation labels.

Although a 2025 study used Vision Transformers (ViTs) for early OA diagnosis, demonstrating encouraging outcomes in identifying long-range dependencies that CNNs frequently overlook, these models' practicality is limited by their significant computational overhead and requirement for huge datasets[31].

All of these studies point to two enduring issues: the need for multi-view or well annotated datasets and the high computational expense of transformer-based or segmentation-driven methods. Current techniques frequently lack efficiency, scalability, and deployability in clinical settings, particularly in settings with limited resources, despite advancements in accuracy. This leads to a research void for accurate and lightweight frameworks that can detect OA from single-view radiographs early with little training overhead, good sensitivity in early-stage classification, and better clinical usability.The Table 1 represents a succinct overview on DL frameworks for the OA knee segmentation and prediction.

Background views

This section discusses about the background views on Swin transformers and Adder networks.

Swin transformers—an overview

A Transformer encoder typically consists of a stack of L identical layers. Each layer is composed of two main components: a Multi-Head Self-Attention (MSA) module and a Multi-Layer Perceptron (MLP). Each sub-

Authors	Techniques used	Accuracy	Computational complexity	Training time	Processing speed	Standard Deviation	Mean	Memory consumption	Power requirement	Robustness	AUC	Overfitting issues
Sindhu et al.	CNN	✓ Average	✗ High	✗ High	✗ low	Not analyzed	Not analyzed	✗ High	✓ Average	✓ Average	✗ less	✗ High
Dalia et al.[20]	VGG16, Resnet	✓ Average	✗ High	✗ High	✗ low	Not analyzed	Not analyzed	✗ High	✗ High	✗ low	✓ Average	✗ High
Shen et al.[21]	CNN, ResNet and DeepLabV3P	✗ Low	✓ Average			✓ Average	✓ Average	✗ High	✓ Average	✓ Average	✗ less	✗ High
Janarthanan et al.[22]	Support Vector Machine (SVM), Naive Bayes (NB), Linear Discriminant Analysis (LDA)	✓ Average	✗ High	✗ High	✗ low	Not analyzed	Not analyzed	✓ Average	✗ High	✗ low	✓ Average	✗ High
Felfeliyan et al.[23]	CycleGAN	✓ Average	✗ High	✗ High	✗ low	Not analyzed	Not analyzed	✗ High	✓ Average	✓ High	✗ less	✗ High
Viekash et al.[24]	CNN	✓ Average	✓ Average	✗ High	✗ low	Not analyzed	Not analyzed	✓ Average	✓ Average		✓ Average	✗ High
Yan et al.[25]	Regression methods and heatmaps	✓ Average	✗ High			Not analyzed	Not analyzed	✗ High	✗ High	✗ low	✓ Average	✗ High
Azcona et al.[26]	Transfer learning	✗ Low	✓ Average	✗ High	✗ low	Not analyzed	Not analyzed	✓ Average	✗ High	✗ low	✓ Average	✗ High
Supatman et al.[27]	DCNN	✓ Average	✗ High	✓ Average	✓ Average	Not analyzed	Not analyzed	✓ Average	✗ High	✓ Average	✓ Average	✗ High
Zhang et al.[28]	CNN	✓ Average	✓ Average	✓ Average	✓ Average	Not analyzed	Not analyzed	✗ High	✗ High	✓ Average	✓ Average	✗ High
Rahman et al.[31]	Multi-view CNN	✓ AboveAverage	✓ AboveAverage	✓ Average	✓ Average	Not analyzed	Not analyzed	✗ High	✓ Average	✓ Average	✓ Average	✗ High
Pan et al.[29]	Hierarchical CNN + Segmentation	✗ High	✗ High	✗ High	✗ Low	✓ Average	✓ Average	✗ High	✗ High	✓ Average	✗ High	✗ High
Chen et al.[30]	Vision Transformer	✗ High	✗ High	✗ High	✗ Low	Not analyzed	Not analyzed	✗ High	✗ High	✗ High	✗ High	✗ High

Table 1. Quicksummary on deep learning frameworks for the OA knee segmentation and prediction.

layer is preceded by a Layer Normalization (LN) operation, and residual connections are added after both the attention and MLP sub-layers.

While conventional Vision Transformers achieve strong performance, they suffer from quadratic computational complexity with respect to the number of image tokens, making them inefficient for dense prediction tasks and high-resolution images. To overcome this, the Swin Transformer[32] introduces two efficient attention mechanisms: Window-based MSA (W-MSA) and Shifted Window MSA (SW-MSA).

- In W-MSA, the input feature map is partitioned into non-overlapping windows of size $M \times M$. Self-attention is then computed only within each local window, reducing computational cost.

- SW-MSA shifts the window partitioning by half the window size between successive layers, enabling cross-window connections and improved long-range modeling.

The computation within a Swin Transformer block can be expressed as:

$$y_2 = W - MSA(LN(y_1)) + y_1 \tag{1}$$

$$y\_out = MLP(LN(y_2)) + y_2 \tag{2}$$

Here, y_1 is the block input, y_2 is the output of the attention module with residual connection, and y_out is the final block output after the MLP sub-layer.

Moreover, Swin transformers exhibits the edge of advantages over the other transformer networks in the computer vision applications. Swin Transformer Module was shown in Fig. 1

Multi-scale attention networks in Swin transformers

The attention mechanism allows networks to focus on informative features while suppressing redundant ones[35]. However, standard attention modules often struggle to jointly capture both local details and global dependencies, since they operate with a fixed receptive field.

The Swin Transformer addresses this by adopting a hierarchical multi-scale representation. Through patch merging, the spatial resolution is progressively reduced while the channel dimension is increased. This design enables the network to extract a rich set of multi-scale features with lower computational overhead.

Each stage of the encoder consists of multiple Swin Transformer blocks that alternate between W-MSA and SW-MSA, thereby integrating both local and long-range feature dependencies. The computation of two successive blocks can be expressed as:

$$y_2 = W - MSA(LN(y_1)) + \ y_1 \tag{3}$$

$$y_1' = MLP(LN(y_2)) + y_2 \tag{4}$$

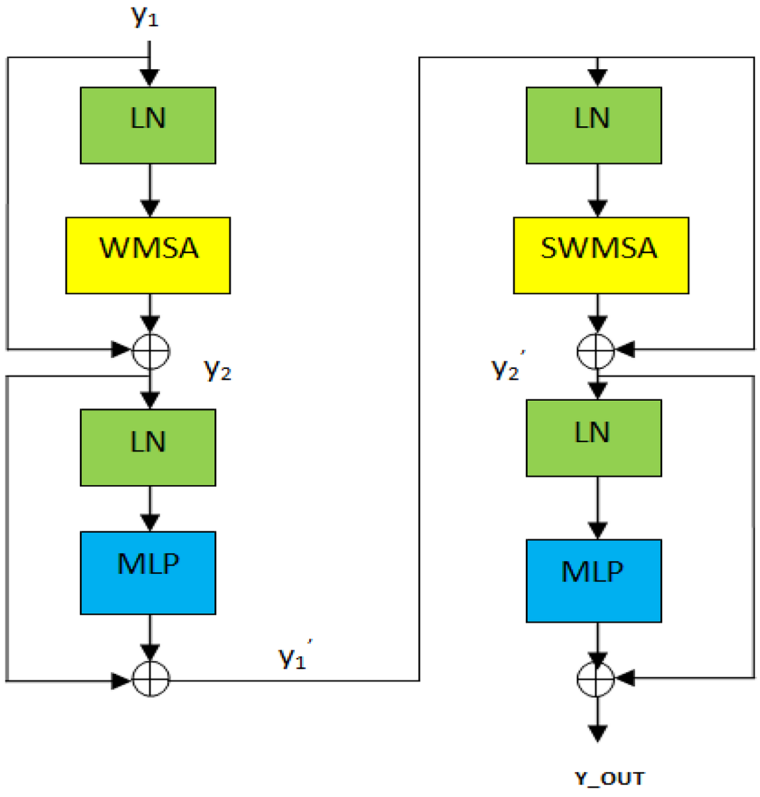

Fig. 1. Swin transformer module.

$$y'_2 = SW - MSA(LN(y'_1)) + y'_1 \quad (5)$$

$$y\_out = MLP(LN(y'_2)) + y'_2 \quad (6)$$

In the above Eqs. (3–4) represent the W-MSA block and Eqs. (5–6) represent the SW-MSA block.Together, they form a multi-scale attention layer, which enables effective modeling of both local and global context.

However, multi-scale attention layers in Swintransfomers are more complex to deploy to obtain the good accuracy.

Integration with fast extreme learning network (FELN)

For classification, a Fast Extreme Learning Network (FELN) is coupled with the Swin Transformer to improve accuracy and efficiency. By employing W-MSA and SW-MSA, multi-scale attention, residual connections, and LN layers to capture hierarchical features, the MST encoder is able to extract both long-range global patterns and fine-grained local patterns from knee X-ray images. This is enhanced by FELN, which functions as a lightweight classifier by initializing hidden layer weights at random and computing output weights in closed form, which significantly cuts down on training time without sacrificing classification performance. This combination makes Swin-O-NETS appropriate for early-stage detection in clinical settings by enabling it to accomplish precise OA severity classification with less computing cost.

Proposed architecture

A redesigned swin transformer block connected in a U-shaped architecture with encoder and decoder in Fig. 3 integrated Fast Extreme Learning networks to obtain improved classification performance is the overall design for the recommended design, as seen in Fig. 2. A skip connection links the encoder and decoder blocks. The suggested transformer block is utilised to efficiently split various Open Access photos based on their semantic content. Secondly, the proposed network incorporates the feature extraction map layer and the multi-headed channel attention maps are designed into the module. The retrieved attributes are ultimately sent to the classification layer, which is constructed using Fast Extreme Learning networks. The section before this one goes into detail on the suggested architecture.

Materials and methods

The Osteoarthritis (OA) Initiative dataset, publicly available via Kaggle[33], served as the foundation for the datasets utilized in this study. This dataset comprises knee X-ray images collected from a large cohort of individuals with varying degrees of osteoarthritis severity, annotated according to the Kellgren–Lawrence (KL) grading system (grades 0–4).The distribution of images across KL grades was carefully considered to mitigate class imbalance, with data augmentation techniques such as rotation, flipping, and scaling applied to increase diversity and improve model generalization. Almost 9786 X-ray pictures were assessed using the KL grading schemes. The various X-ray image groups according to severity ratings are shown in Table 2. Every image has a size of 224×224. Due to the extreme imbalance in the data, it has been divided into train and test (70:30) based on the quantity of samples for each category that are available.

Data pre-processing

Using the medical preprocessing technique, noise and low-quality pixels that obstruct the diagnosis of OA illnesses are reduced. The Pixel Intensive Testing method has been used to remove the noise and erroneous pixels from the knee images. Additionally, image histogram techniques are utilized to enhance image quality since they work better on a range of images.

Data augmentation process

The recommended architectural design uses the image augmentation technique after preprocessing the input photos. When there is a limited amount of labelled data available, Neural Networks might cause overfitting issues. This challenge can be solved most effectively and efficiently via data augmentation. As the data enhancement procedure progresses, every picture undergoes several modifications to produce the enormous quantity of fresh examples of training images that have been rectified. For an effective data augmentation, affine transformation is used, as covered in[34]. Affine transformation methods like as there is the usage of rotation, scaling, and translation. To alleviate over fitting concerns, this step is recommended because the majority of training picture samples acquired during the augmentation procedure display correlation.

Network model and design

The basic segmentation framework for the suggested design is according to the U-Net design, as depicted in Fig. 3, as was covered in Section "Multi-scale attention networks in Swin transformers". It is possible to think of this design in the form of U-Net interface that combines the intended Swin transformer models. The four components of the proposed network are skip connections, encoders with Swin Transformer activated, decoders, and patch extraction. The suggested model's capacity to increase U-nets' expressive performance and extract richer, deeper features is one of its main advantages.

To extract the deeper features, this research incorporates the Multi-Headed Channel Self Attention layers (MHCSA) in the place of standalone self-attention maps interfaced with MLP mechanism. This modified Swin architecture is integrated both in encoder and decoder structures. In encoder, these structures are employed to extract rid the channel based contextual features whereas the low-level features are then fused by the proposed structure which can aid in the increased performance and classification accuracy. The detailed description of each and every layer is detailed as follows:

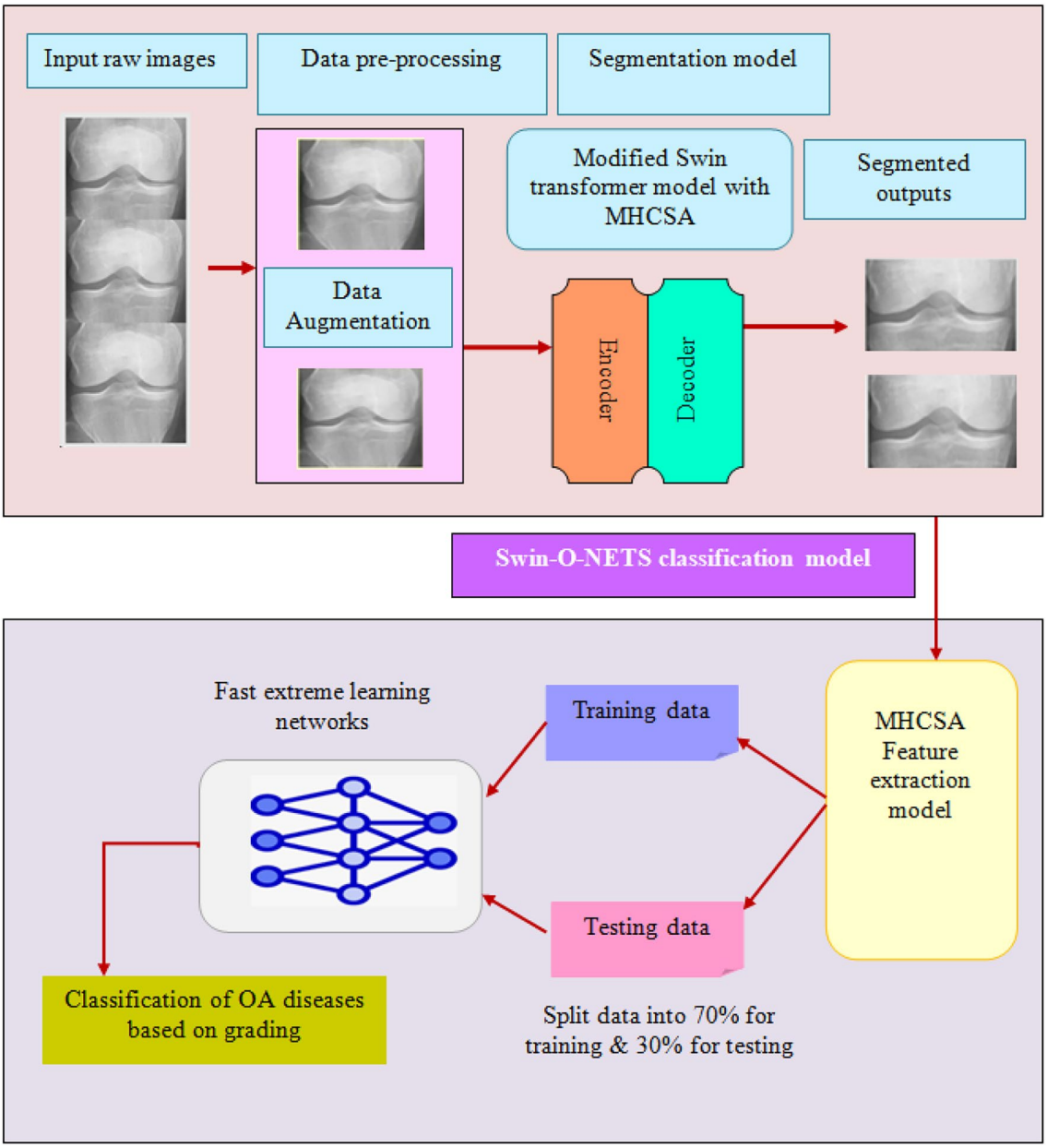

Fig. 2. Proposed Framework for Swin-O-NETS deployed for classification of OA datasets with severity grading.

Type of X-rays images	Description of the images	Number of images
Grade-0	Healthy knee image	647
Grade-1	Doubtful joint narrowing with possible Osteophytic lipping	367
Grade-2	Definite presence of osteophytes and possible joint space narrowing	387
Grade-3	Multiple osteophytes, definite joint space narrowing, with mild sclerosis	234
Grade-4	Large osteophytes, significant joint narrowing, and severe sclerosis	412
	Total images	2047 Images

Table 2. Illustration of the X-ray images used for evaluating the proposed networks.

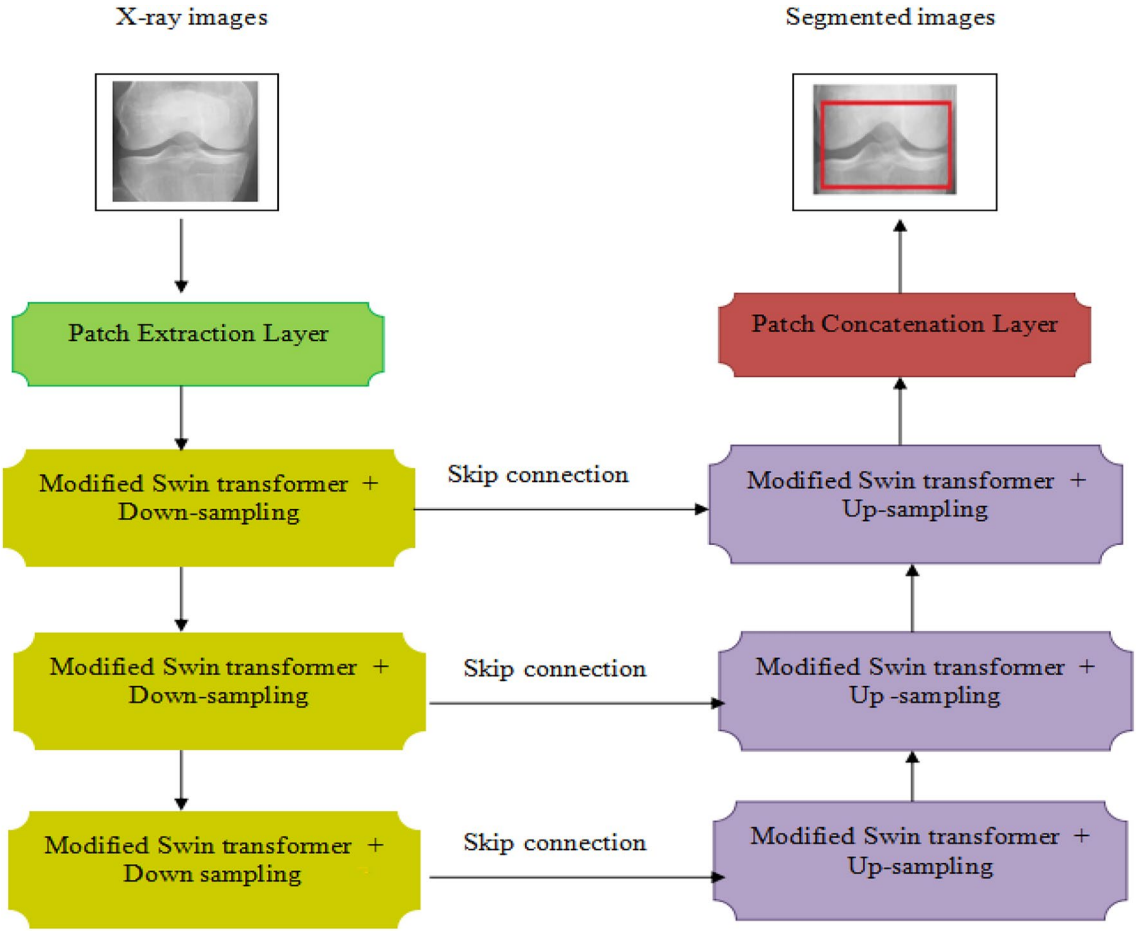

Fig. 3. U-Net based modified Swin transformer method used for the segmentation.

Multi-headed channel self attention layers

These attention maps are considered as the variant of self –attention maps which combines the Channel attention with Self –attention maps to enrich the feature extraction for which the segmentation performance can readily increase. The Channel attention maps are used to obtain most significant features surpassing the useless features by performing the two multiplication operations, softmax, and an addition operation.The general mathematical procedures performed in channel self-attention maps are given in Eq. (7). The three vectors query (Q), key (K), and value (V) spacesare then created for each input sequence X using the retrieved characteristics, and these are fed to the self-attention layers.

$$\mathrm{Q} = \mathrm{XW\_Q}, \mathrm{K} = \mathrm{XW\_K}, \mathrm{V} = \mathrm{XW\_V}$$

The raw channel self-attention mapping is given as:

$$\mathrm{CSA(X)} = \mathrm{Softmax(I1(X) \cdot I2(X)\text{^}T)I3(X)} \tag{7}$$

where I1, I2, I3 are learned projections corresponding to Q, K, V.

The query is mapped with the set of key pairs using the scaled dot-product attention

$$\mathrm{CSA(Q, K, V)} = \mathrm{Softmax((QK\text{^}T)/sqrt(d\_k))V} \tag{8}$$

- Q, K∈R^(C×d_k), V∈R^(C×d_v)
- d_k is the key/query dimension
- Scaling by sqrt(d_k) ensures stable gradients

The following Eq. (8) is repeated 'n' how long it takes to create values matrices, keys, and queries in order to further refine the useful information based on the pictures. A mathematical expression for the MHCSA framework is

$$\mathrm{MHCSA\ (Q, K, V) = Concat\ (CSA1\ (Q, K, V)\ ,\ CSA2\ (Q, K, V)\ ,\ \ldots\ ,\ CSAh\ (Q, K, V))\ W\_O} \tag{9}$$

where h is the number of heads, and W_O is the final output projection.

These MHCSA is incorporated in Swin transformers to extract the more features which.

aids for the segmentation process. As discussed in the Section "Background views", Eqs. (3–6) are modified in accordance to the above Eq. (9)

$$y_2 = W - MHCSA(LN(y_1)) + y_1 \tag{10}$$

$$y_1' = MLP(LN(y_2)) + y_2 \tag{11}$$

$$y_2' = SW - MHCSA(LN(y_1')) + y_1' \tag{12}$$

$$y\_out = MLP(LN(y_2')) + y_2' \tag{13}$$

Encoder design

The general architecture of the suggested the simulation is predicated upon a U-shaped framework, as was previously mentioned. Modified Swin transformers employ the MHCSA network for the encoder, followed by layers of downsampling. Input pre-processed medial pictures are first separated through non-overlapping regions of dimension Ls X Hs, wherein s is the region's dimension, as illustrated in Fig. 3. The overhead-free patches are formed using the basic convolutional layers. The upgraded Swin transformers get these patches, and then the downsampling procedure is applied to generate additional hierarchical features. As a result, each post-downsampling stage's output dimension will increase. The suggested encoder performs a residual skip procedure via an up-sampling action as ψ and sets the encoder outcome to the identical dimension as the feature map on the left side of the structure as E(s). As a result, the relationship between each encoder block is expressed as Encoder stage outputs

$$E(s) = E(M)\ \psi\ F(d) \tag{14}$$

In the first stage, proposed Encoder's outputs are mathematically expressed using Eq. (15)

$$E(M1) = (y1 * y2)\,\psi F(d) \tag{15}$$

The current stage of the second stage receives its input from the first encoder stage's output

$$E(M2) = (E(M1)\ \ \mathrm{Concat}\,(y1 * y2))\,\psi F(d) \tag{16}$$

Third-stage current is fed into the current by the output of the second encoder stage

$$E(M3) = (E(M2)\ \ \mathrm{Concat}\,(y1 * y2))\,\psi\ F(d) \tag{17}$$

After combining each of these encoder stage steps, the final concatenated outputs are created as

$$E(e) = \sum\nolimits_{i=1^3} E(M(i)\ \psi\ F(d(i))) \tag{18}$$

where E(M), is the generating value associated with each encoder level individually. The resultants of the downsampling phase that are produced both before and after the skip connections are represented by F(dp).

Decoder model

As Fig. 4 shows, the decoder is mostly composed of three phases. In contrast to the earlier U-Net and its version, the suggested model incorporates MHCSA block in addition to skip connection and up-sampling at each level. Specifically, the encoder's stage 4 output serves as the decoder's initial input. The relevant skip connection feature maps from the corresponding encoder phase are concatenated first, the input features in each decoder step undergo an upsampling by two processes. After that, the output is sent to the Modified Multi-Head Attention block with Residual Connections and Transformer. The recommended transformer block in decoders has the following advantages:

a. Enables the decoder to fully utilise the encoder's features and upsampling
b. Creating long-range dependencies improves decoding performances.

Following the aforementioned three processes, a final result featuring a resolution of L/4 × H/4 is produced. Many shallow features are lost when using a 4 × upsampling operator directly, thus we downscale the input picture by merging the two blocks to obtain low level features with resolutions of L × B and L/2 × B/2. Each block is made up of modified swin transformers and an upsampling operation. By using a skip connection, all of these output attributes will be exploited to produce smoother final photos. The following is the mathematical expression for each decoder stage

$$\text{Decoder Stage outputs } D(s) = U(M)\,\psi F(U) \tag{19}$$

First, the recommended decoder's outputs are expressed mathematically as

$$U(M1) = (y1 * y2)\,\emptyset\ F(u) \tag{20}$$

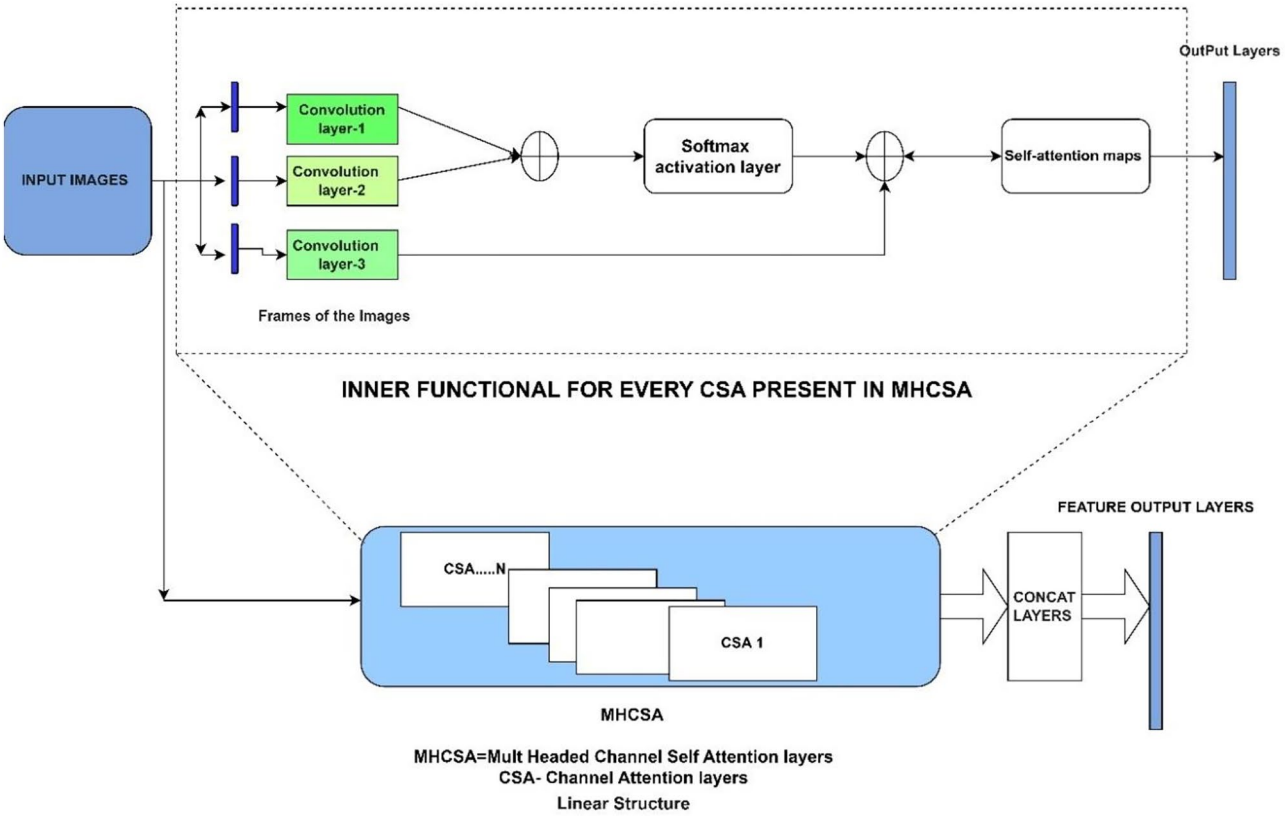

Fig. 4. Proposed MHCSA architecture incorporated in Swin transformers.

First decoder stage output serves as the input for the current step of the second stage

$$\mathrm{U}\,(\mathrm{M2}) = (\mathrm{U}\,(\mathrm{M1})\ \ \mathrm{Concat}\,(\mathrm{y1} * \mathrm{y2}))\,\emptyset\ \mathrm{F}\,(\mathrm{u}) \tag{21}$$

In the third stage, the input of the current stage is the output of the second decoder stage

$$\mathrm{U}\,(\mathrm{M3}) = (\mathrm{U}\,(\mathrm{M2})\ \ \mathrm{Concat}\,(\mathrm{y1} * \mathrm{y2}))\,\emptyset\ \mathrm{F}\,(\mathrm{u}) \tag{22}$$

After combining each of these encoder stage steps, the final concatenated outputs are created as

$$\mathrm{U}\,(\mathrm{e}) = \sum\nolimits_{\mathrm{i}=1^{5}} \mathrm{U}(\mathrm{M}\,(\mathrm{i})\,\emptyset\ \mathrm{F}\,(\mathrm{u}\,(\mathrm{i})) \tag{23}$$

In each encoder step, u(M) represents the output function, which is determined by Eq. (20). The outputs of the downsampling step produced both prior to and after to the skip connections are denoted by F(d). Similar to U-Net, skip connections are used to fuse multiscale features from the encoder with up-sample features from the decoder.

Feature extraction layers

Once the segmented images are obtained, the last hurdle is to effectively extract the features. To be able to record the extended relationships and dependencies among the features, again MHCSA is adopted for the better feature extraction which can integrate the multi-scale features which can further improvises the classification layers.

Fast extreme learning networks

In order to speed computation while preserving a high degree of accuracy in the classification of skin cancers, this study employs unique fast neural networks created using the ELM principle. G.B.Huang[34] proposed the idea of extreme learning machines, and this is how the proposed model classifies grades quickly and accurately. Single hidden layers are used in this kind of neural network; adjusting them is necessary.

The kernel function is used by ELM to deliver better efficiency and effectiveness. The ELM's principal benefits are its low training error and high approximation quality. For classification and classification values, ELM is useful since it uses non-zero activation functions and automatically adjusts weight biases.

Although the activation function of the output layer is straight, the hidden layer's' L' neurons need to interact with an extremely distinctive one (such as the sigmoid function) in order to achieve optimal performance.

If you're using ELM, tuning hidden layers is optional. The ELM protocol does not require the hidden layer to be tweaked. The buried layer's loads (including the bias loads) are decided upon at will. Furthermore, is it possible to generate the concealed neurons' settings at random beforehand, but hidden nodes themselves are not necessary for the network to function.

In other words, prior to the training set's data being processed. For an ELM with one hidden layer, the system yield is given by Eq. (24).

$$\mathrm{f_L}\left(\mathrm{x}\right) = \sum\nolimits_{\mathrm{i}=1^{\mathrm{L}}} \beta_{\mathrm{i}} \mathrm{h_i}\left(\mathrm{x}\right) = \mathrm{h}\left(\mathrm{x}\right)\ \beta \tag{24}$$

here, x stands for input features of the encoder-decoder.

The output weight vector is displayed in the manner described below:

$$\beta = \left[\beta_1, \beta_2, \ldots \beta_{\mathrm{L}}\right]^{\mathrm{T}} \tag{25}$$

h(x) is the output of the hidden layer, and its equation may be seen below:

$$\mathrm{h}\left(\mathrm{x}\right) = \left[\mathrm{h}_1\left(\mathrm{x}\right),\ \mathrm{h}_2\left(\mathrm{x}\right), \ldots \mathrm{h}_{\mathrm{L}}\left(\mathrm{x}\right)\right] \tag{26}$$

The hidden layers are expressed through Eq. (28) to produce the O, additionally referred to by the term target vector.

$$\mathrm{H1} = \begin{bmatrix} \mathrm{h}\left(\mathrm{x}_1\right) \\ \mathrm{h}\left(\mathrm{x}_2\right) \\ \vdots \\ \mathrm{h}\left(\mathrm{x}_{\mathrm{N}}\right) \end{bmatrix} \tag{27}$$

The essential architecture of the ELM makes use of minimal non-linear least square techniques as shown by Eq. (7)

$$\beta' = \mathrm{H} * \mathrm{O} = \mathrm{H}^{\mathrm{T}}\left(\mathrm{HH}^{\mathrm{T}}\right)^{-1}\mathrm{O} \tag{28}$$

where H is the Moore–Penrose generalised inverse of H*→.

Another form of the above equation is as follows.

$$\beta' = \mathrm{H}^{\mathrm{T}}\left(1/\mathrm{c} * \mathrm{HH}^{\mathrm{T}}\right)^{-1}\mathrm{O} \tag{29}$$

Therefore, the above equation can be used to get the output function.

$$\mathrm{f_L}\left(\mathrm{x}\right) = \mathrm{h}\left(\mathrm{x}\right)\ \beta = \mathrm{h}\left(\mathrm{x}\right)\ \ \mathrm{H}^{\mathrm{T}}\left(1/\mathrm{c} * \mathrm{HH}^{\mathrm{T}}\right)^{-1}\mathrm{O} \tag{30}$$

Ultimately, the suggested classification layer incorporates Categorical cross-entropy to improve performance in grading the picture datasets. Algorithm-1 depicts the operating system of the suggested approach.

Step	
1	Input : Pre-processed Images
2	Outputs : Knee X-ray Images
3	Segmentation of the Images using Modified Swin Transformer
4	Feature extraction from the Segmented Images using MHCSA
5	OA based classification using the Equation(28)
6	If ($f_L(x) = = 1$)
7	//Normal Image is determined
8	Else If ($f_L(x) = = 2$)
9	//Grade-1 is determined
10	Else If ($f_L(x) = = 3$)
11	//Grade-2 is detected
12	Else If ($f_L(x) = = 4$)
13	//Grade-3 is detected
14	Else If($f_L(x) = = 5$)
15	//Grade-4 is detected
16	Else
17	Go to Step 5
18	End
19	End
20	End
21	End

Algorithm 1.

Experimental results and analysis

Experimentation process

The experimentation of the complete architecture was implemented with Python 3.10, Tensorflow version2.2 and keras as front end in Google Co-lab Environment. To optimize the final classification layers, the ADAM optimizer starts with a learning rate of 0.00001. A batch size of 32 and 200 epochs, respectively, were used to train the framework. Following the evaluation of each epoch validation set, the cross-entropy is used in loss estimates. Table 3. lists the hyper parameters used for the training the network.

Evaluation metrics

The predictions will be divided into several classes by a DL model before being assessed. The measurement of a detected image's accuracy is called True Positive (TP). A picture that is identified as positive even when it is negative is called a False Positive (FP). When an image is labelled as True Negative (TN), it indicates that it is indeed negative. When an image is classified as false negative (FN), it is actually positive. The performance of the proposed Swin-O-NETS framework was evaluated using multiple metrics to provide a comprehensive assessment. Accuracy represents the proportion of correctly classified knee X-ray images across all KL grades, indicating the overall prediction capability of the model. Precision measures the fraction of correctly predicted instances among all predicted instances for each class, reflecting reliability in identifying specific OA severity levels. Recall (Sensitivity) quantifies the model's ability to correctly detect true positive cases, emphasizing effectiveness in recognizing actual OA instances. The F1-score, which combines precision and recall, offers a

Sl.no	Hyper-parameters	Specification
1	Batch size	32
2	Epochs	200
3	Optimizer	ADAM
4	Learning rate	0.00001
5	Loss function	Cross entropy

Table 3. Hyperparameters used for training the network.

SL.NO	Performance metrics	Mathematical expression
01	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
02	Sensitivity or recall	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}\ \frac{TP}{TP+TN} \times 100$
03	Specificity	$\frac{TN}{TN+FP}$
04	Precision	$\frac{TP}{TP+FP}$
05	F1-score	$2.\frac{Precision*Recall}{Precision+Recall}$
06	Mathew correlation co-efficient (MCC)	$(TP+TN)/\sqrt{(TP+FP)+(TN+FN)+(TN*FP)+(TP*FN)}$

Table 4. The suggested approach is evaluated using indicators of performance.

Algorithms	Performance metrics (%)					
	Accuracy	Precision	Recall	Specificity	F1-score	MCC
CNN	84.3	83.2	82.3	81.4	82.4	0.82
VGG-19	74.3	73.2	72.1	70.4	72.5	0.72
DenseNETS	93.2	90.5	90.3	89.5	90.4	0.88
RESNETS-34	91.4	91	90.4	90.2	90.7	0.90
Ensemble methods	95.4	94.3	94.2	94	94.3	0.94
Resnets + CBAM	93.2	93.4	92.4	93.0	92.9	0.92
Proposed model	99.5	98.9	98.7	98.4	98.6	0.98

Table 5. Effectiveness measures of various DL techniques in identifying Grade-0 severity.

balanced metric particularly valuable in the presence of class imbalance. Area Under the ROC Curve (AUC) demonstrates the model's discrimination ability across different severity classes. Additionally, computational complexity, memory usage, and training time were analyzed to evaluate the framework's efficiency and practical applicability.

The proposed Swin-O-NETS achieves superior performance across these metrics due to the hierarchical feature extraction capability of the Modified Swin Transformer, which captures both local and global structural patterns in knee X-rays, and the Fast Extreme Learning Network, which provides fast and accurate classification with low computational overhead. The inclusion of multi-scale attention further enhances the detection of subtle joint changes, enabling early-stage OA grading and contributing to improved precision, recall, and overall reliability compared to existing approaches.

Metrics for evaluation such as MCC, F1-score, recall, specificity, accuracy, and precision are computed using the mathematical formula presented in Table 4 in relation to the data.

Besides, the Receiver Operating Characteristics Curve (ROC) and Confusion Matrix for the different model are calculated and analysed.

Results and discussion

Using the metrics listed in Table 4, the proposed model's classification performance was compared with those of the other cutting-edge DL models in this section. DL architectures were taken into consideration for assessing the classification performance. It was crucial to remember that each framework is trained with identical datasets and the same experimental conditions as those mentioned in Section "Proposed architecture". Test data is subsequently used to assess and validate the trained samples. For every scenario, 70% of the datasets were used for training, and 30% were used for testing. The average distribution of the datasets used for testing and training is shown in confusion matrix. The way in which the suggested methodology in categorizing the various strictness classes of the intended photos is shown in Tables 5, 6, 7, 8 and 9. It is accomplished by comparing it with other DL models.

The findings unequivocally show that the suggested Swin-O-NETS framework performs better at every level of knee OA severity. Although the accuracy, precision, recall, F1-score, specificity, and MCC of hybrid ResNet models combined with CBAM and ensemble learning techniques are excellent, their performance slightly degrades when classifying different KL grades. By efficiently recording patch-level information and both local and global structural patterns, the Modified Swin Transformer with Multi-Headed Channel Attention, on the other hand, continuously provides exceptional and consistent performance across all severity levels. The suggested transformer network outperforms current architectures due to its capacity to extract rich, hierarchical features, demonstrating its resilience and appropriateness for automated OA severity categorization. Figures 5, 6, 7, 8, 9 and 10 show the confusion matrix of the different learning algorithms along with the proposed model. Out of 647 garde-0 images, CNN has detected 615 Images, VGG-19 has detected 620 images while RESNETS, DENSENETS and Hybrid RESNETS + CBAM has detected 627,632, 639 respectively. But the proposed model has detected 645 images correctly making it superiority over the other deep learning algorithms.

Algorithms	Performance metrics					
	Accuracy	Precision	Recall	Specificity	F1-score	MCC
CNN	82.3	81.2	80.3	79.5	81.2	0.81
VGG-19	77.2	76.2	74.1	70.4	70.5	0.72
DenseNETS	92.4	92.4	90.3	89.5	90.4	0.88
RESNETS-34	92.2	92.1	90.4	90.2	90.7	0.90
Ensemble methods	96.1	95.4	94.2	94	94.3	0.94
Resnets + CBAM	93.5	92.9	92.4	93.0	92.9	0.92
Proposed model	99.67	99.0	98.9	98.7	98.9	0.98

Table 6. Grade-1 severity detection using performance metrics of various DL methods.

Algorithms	Performance metrics					
	Accuracy	Precision	Recall	Specificity	F1-score	MCC
CNN	84.3	83.2	82.3	81.4	82.4	0.82
VGG-19	74.3	73.2	72.1	70.4	72.5	0.72
DenseNETS	93.2	90.5	90.3	89.5	90.4	0.88
RESNETS-34	91.4	91	90.4	90.2	90.7	0.90
Ensemble methods	95.4	94.3	94.2	94	94.3	0.94
Resnets + CBAM	93.2	93.4	92.4	93.0	92.9	0.92
Proposed model	99.5	98.9	98.7	98.4	98.6	0.98

Table 7. Effectiveness measures of various DL techniques in identifying Grade-2 severity.

Algorithms	Performance metrics					
	Accuracy	Precision	Recall	Specificity	F1-score	MCC
CNN	84.3	83.2	82.3	81.4	82.4	0.82
VGG-19	74.3	73.2	72.1	70.4	72.5	0.72
DenseNETS	93.2	90.5	90.3	89.5	90.4	0.88
RESNETS-34	91.4	91	90.4	90.2	90.7	0.90
Ensemble methods	95.4	94.3	94.2	94	94.3	0.94
Resnets + CBAM	93.2	93.4	92.4	93.0	92.9	0.92
Proposed model	99.5	98.9	98.7	98.4	98.6	0.98

Table 8. Grade 3 severity detection using performance metrics of various DL methods.

Algorithms	Performance metrics					
	Accuracy	Precision	Recall	Specificity	F1-score	MCC
CNN	83.5	82.1	81.0	80.2	81.5	0.81
VGG-19	73.0	72.0	70.5	69.8	71.2	0.70
DenseNETS	92.8	91.0	90.2	89.5	90.6	0.88
RESNETS-34	91.2	91.0	90.0	90.1	90.5	0.89
Ensemble methods	95.0	94.0	93.8	93.5	93.9	0.93
Resnets + CBAM	93.0	93.2	92.2	92.5	92.7	0.92
Proposed model	99.5	99.0	98.8	98.5	98.9	0.98

Table 9. Grade 4 severity detection using performance metrics of various DL methods.

The confusion matrix in Figs. 5, 6, 7, 8, 9 and 10 illustrate the comparable type of detection performance that is also observed for grades 0, 1, 2, 3, and 4, correspondingly.

The knee osteoarthritis classification performance of several models is compared using the ROC curve in Fig. 11 The suggested Swin-O-NETS outperforms CNN, ResNet, DenseNet, and ensemble models in terms of accuracy, with the greatest AUC of 0.9838. According to the results, Swin-O-NETS substantially outperforms conventional topologies and offers dependable and strong severity grading.

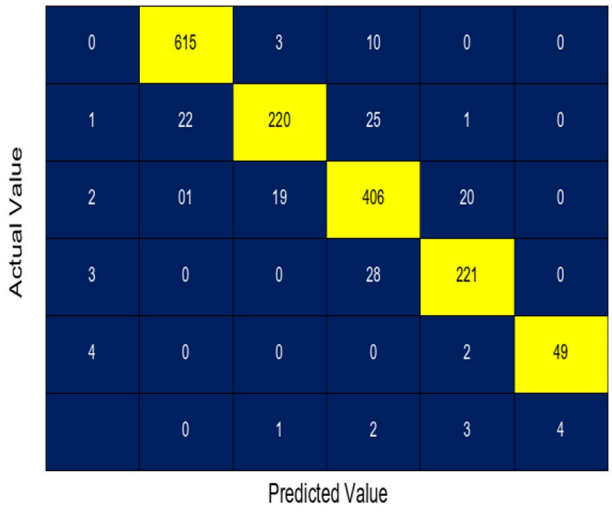

Fig. 5. CNN.

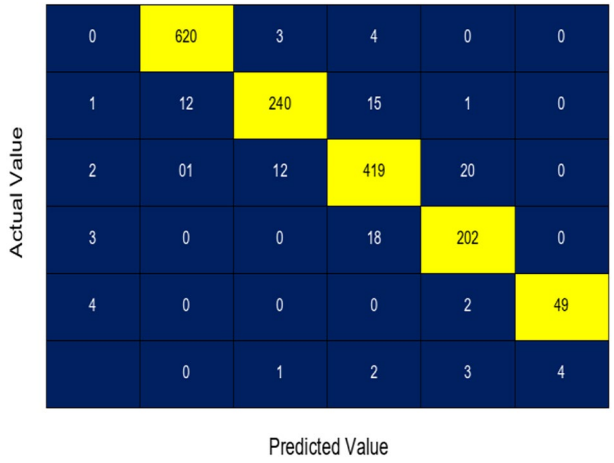

Fig. 6. VGG-19.

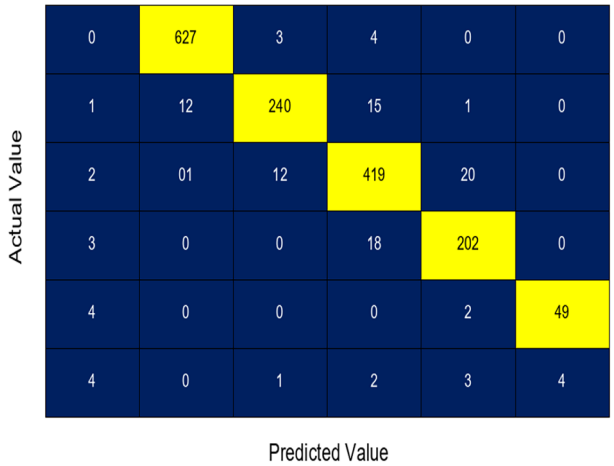

Fig. 7. RESNETS-34.

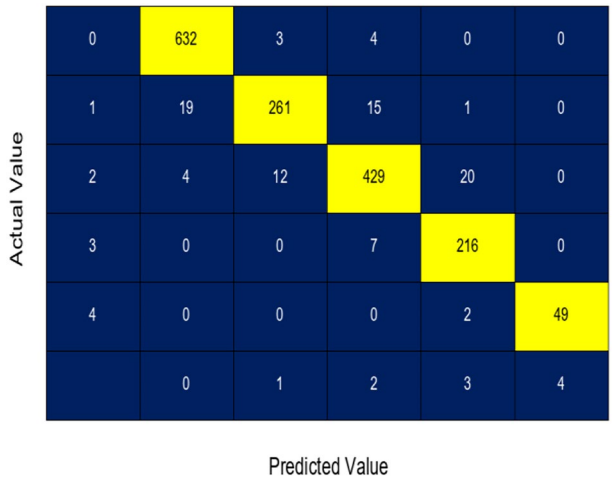

Fig. 8. DENSENETS-121.

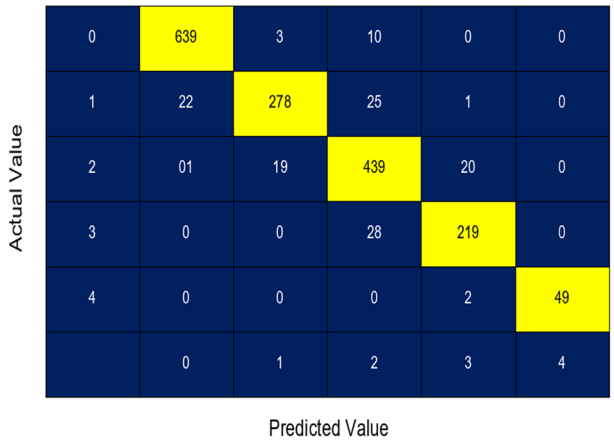

Fig. 9. Hybrid RESNETS + CBAM.

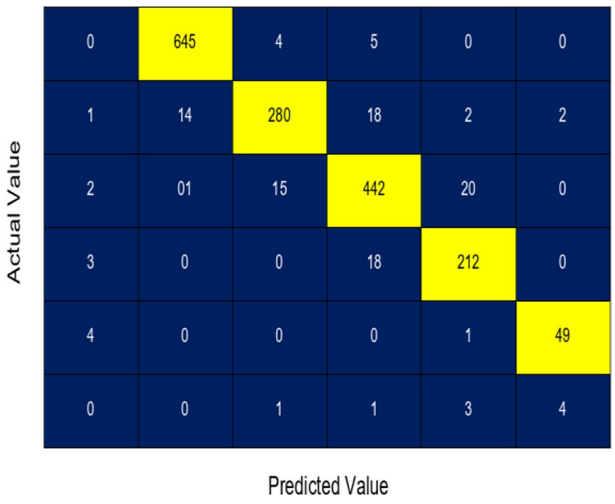

Fig. 10. SWIN-O-NETS.

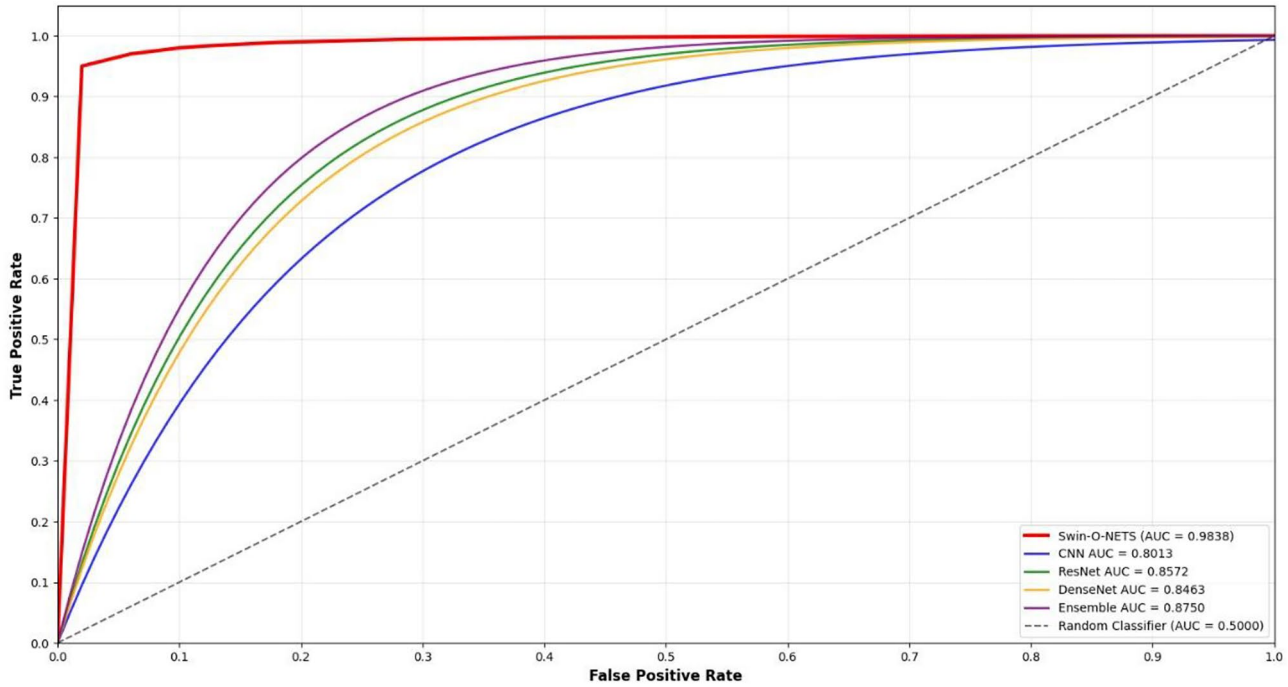

Fig. 11. ROC curves for Swin-O-NETS vs comparative models.

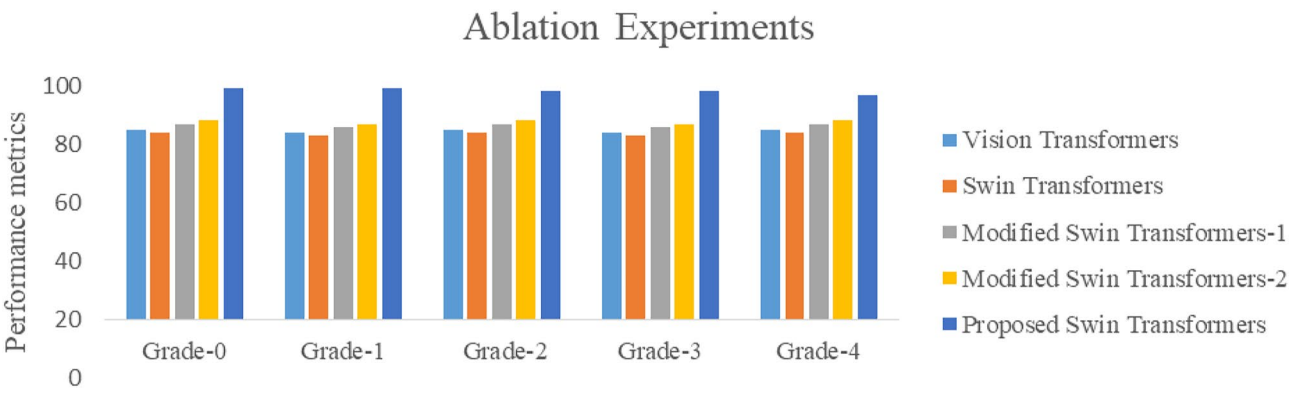

Fig. 12. Ablation experiments for the different vision and Swin transformers used for classifying the OA images with severity grades.

Ablation experiments

In this part, ablation experimentation is got to prove the efficacy of the suggested Swin transformers and Multi-headed Channel Attention layers in classifying the different severity grades. The examination of throughout the training period, every one strategy loss varies to evaluate how well it works, culminates in the figure that is displayed. To prove the excellence of the proposed SWIN models, ablation experimentation is also carried with the other existing transformer models. The x-axis represents the KL grades, while the y-axis shows the performance metrics (%). The comparison includes Vision Transformers, Swin Transformers(Baseline), Modified Swin Transformer-1 (Baseline Swin Transformer + MHCSA), Modified Swin Transformer-2 (Swin + MHCSA + Multi-Level Feature Fusion), and the Proposed Swin Transformers (Swin + MHCSA + Multi-Level Feature Fusion + FELN Classifier). As shown in Fig. 12 module of proposed SWIN transformer networks could significantly improve the segmentation part of the model and it converge to a lower loss value of training process.

The suggested model is more effective than other transformer algorithms already in use when it comes to identifying OA disorders according to severity classes, as demonstrated by the ablation procedure. Figure 12 shows that the ablation experiments carried out for the transformers models used for the image processing applications. Additionally, ablation experiments had been carried out and contrasted to the suggested approach and other cutting-edge learning models. The ablation result analysis for the various methods used to categories the various grades of OA illnesses is displayed in Fig. 13.

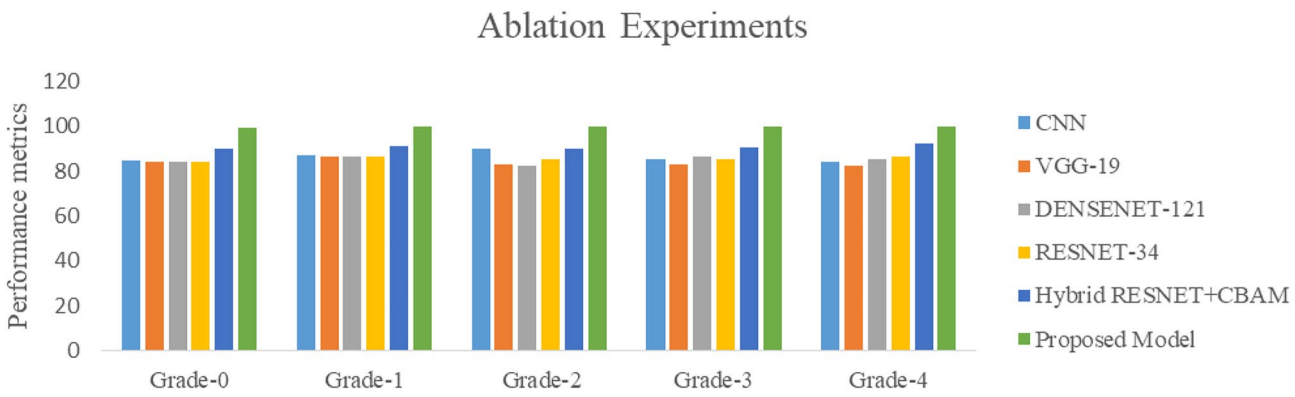

Fig. 13. Ablation result analysis for the different algorithms in classifying the different grades of the OA diseases.

The results of the evaluations show that the suggested model has performed better than other transformer networks and other innovative models.

Conclusion and future discussion

This study suggests an innovative Swin Transformer combined with Fast Extreme learning networks for an ordinal classification method of knee osteoarthritis X-ray grading. Novel multi-headed channel self-attention maps are integrated to optimize the performance and achieve the best classification performance for the Swin Transformers. The Kaggle OA datasets with varying KL gradings are used for the comprehensive testing. Some indicators of performance, that include recall, accuracy, precision, and F1-score are computed and contrasted with another cutting-edge DL methodologies. The approach supports early and accurate OA detection by effectively capturing both local and global joint patterns through hierarchical feature extraction and multi-headed attention.

Despite its benefits, Swin-O-NETS may have drawbacks, including as susceptibility to changes in image quality and the computational needs of multi-scale attention. Future research might concentrate on validate the model on bigger, multi-center datasets, create lightweight, real-time deployable versions for clinical settings, and integrate multimodal imaging data. A promising method for automated, early OA evaluation, these modifications will further improve the system's interpretability, efficiency, and application.

Data availability

The data used in this study are publicly available from the Osteoarthritis Initiative (OAI) database at https://nda.nih.gov/oai and dataset from kaggle at https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity.

Code availability

The code used to train and evaluate the models CNN, VGG19, ResNet34, DenseNet121, Hybrid RESNET + CBAM, SWIN-O-NET are publicly available at: https://github.com/sudha120/Knee_OA.

Received: 17 September 2024; Accepted: 10 March 2026
Published online: 30 March 2026

References

1. Badshah, Y. et al. Genetic markers of osteoarthritis: Early diagnosis in susceptible Pakistani population. *J. Orthop. Surg. Res.* **16**(1), 1–8 (2021).
2. Das, S. K. & Farooqi, A. Osteoarthritis,BestPract. *Res. Clin. Rheumatol.* **22**(4), 657–675 (2008).
3. Gornale, S. S., Patravali, P. U. & Manza, R. R. Detection of osteoarthritis using knee X-ray image analyses: A machine vision based approach. *Int. J. Comput. Appl.* **145**(1), 20–26 (2016).
4. Tas, N. P. et al. A ASNET: A novel AI framework for accurate ankylosing spondylitis diagnosis from MRI. *Biomedicines* **11**, 2441. https://doi.org/10.3390/biomedicines11092441 (2023).
5. Kaya, O. & Taşcı, B. A pyramid deep feature extraction model for the automatic classification of upper extremity fractures. *Diagnostics* **13**(21), 3317 (2023).
6. Pal, C. P. et al. Epidemiology of knee osteoarthritis in India and related factors. *Indian J. Orthopaedics* **50**(5), 518–522 (2016).
7. Teoh, Y. X. et al. Discovering knee osteoarthritis imaging features for diagnosis and prognosis: Review of manual imaging grading and machine learning approaches. *J. Healthc. Eng.* **2022**, 1–19 (2022).
8. Kellgren, J. H. & Lawrence, J. S. Radiological assessment of osteoarthrosis. *Ann. Rheum. Dis.* **16**(4), 494–502 (1957).
9. Saini, D., Chand, T., Chouhan, D. K. & Prakash, M. A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on X-ray images. *Biocybern. Biomed. Eng.* **41**(2), 419–444 (2021).
10. Anifah, L., Purnama, I. K. E., Hariadi, M. & Purnomo, M. H. Osteoarthritis classification using self organizing map based on Gabor kernel and contrast-limited adaptive histogram equalization. *Open Biomed. Eng. J.* **7**(1), 18–28 (2013).
11. Brahim, A. et al. A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis initiative. *Comput. Med. Imaging Graph.* **73**, 11–18 (2019).

12. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778 (2016)
13. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv:1409.1556.
14. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708 (2017).
15. Chaves, E. et al. Evaluation of transfer learning of pre-trained CNNs applied to breast cancer detection on infrared images. *Appl. Opt.* **59**(17), 23 (2020).
16. Yong, C. W. et al. Knee osteoarthritis severity classification with ordinal regression module. *Multim. Tools Appl.* **81**, 41497–41509 (2021).
17. Wu, X. et al. A novel centralized federated deep fuzzy neural network with multi-objectives neural architecture search for epistatic detection. *IEEE Trans. Fuzzy Syst.* **33**(1), 94–107. https://doi.org/10.1109/TFUZZ.2024.3369944 (2025).
18. Shoaib, M. A. et al. Comparative studies of deep learning segmentation models for left ventricle segmentation. *Front. Public Health* **10**, 981019. https://doi.org/10.3389/fpubh.2022.981019 (2022).
19. Shoaib, M. A., Lai, K. W., Chuah, J. H., Hum, Y. C., Ali, R., Dhanalakshmi, S., & Wu, X. Comparative studies of deep learning segmentation models for left ventricle segmentation. *Front Public Health* **10**, 981019(2022).
20. Dalia, Y., Bharath, A., Mayya, V., & Kamath, S. S. DeepOA: Clinical decision support system for early detection and severity grading of knee osteoarthritis, in *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 250–255 (Chennai, India, 2021). https://doi.org/10.1109/ICCCSP52374.2021.9465522.
21. Shen, Z., Laredo, J. D., Lomenie, N., & Chappard, C. Deep learning on knee CT scans from osteoarthritis patients for joint space assessment, in *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 348–353 (Dijon, France, 2022). https://doi.org/10.1109/SITIS57111.2022.00059
22. Janarthanan, V., Pradeep, S., Vidhya, K., Bhosale, S. A., & Kumar, A. Hybrid AI model for arthritis prediction from medical image, in *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, 1476–1482 (Trichy, India, 2022). https://doi.org/10.1109/ICOSEC54921.2022.9952029.
23. Felfeliyan, B., Hareendranathan, A., Kuntze, G., Jaremko, J., & Ronsky, J. MRI knee domain translation for unsupervised segmentation By CycleGAN (data from Osteoarthritis initiative (OAI)), in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 4052–4055 (Mexico, 2021) https://doi.org/10.1109/EMBC46164.2021.9629705.
24. Viekash, V. K., Arun, P., Manimozhi, S., Nagotanekar, G. D., & Deenadayalan, E. Deep learning based muscle intent classification in continuous passive motion machine for knee osteoarthritis rehabilitation, in *2021 IEEE Madras Section Conference (MASCON)*, 1–8 (Chennai, India, 2021. https://doi.org/10.1109/MASCON51689.2021.9563370.
25. Yan, S., Ramazanian, T., Chaudhary, V., & Kremers, H. M. Deep learning method for hip knee ankle angle prediction on postoperative full-limb radiographs of total knee arthroplasty patients, in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 5070–5073 (Glasgow, Scotland, United Kingdom, 2022). https://doi.org/10.1109/EMBC48229.2022.9870936.
26. Azcona, D., McGuinness, K., & Smeaton, A. F. A comparative study of existing and new deep learning methods for detecting knee injuries using the mrnet dataset, in *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 149–155 (Valencia, Spain, 2020). https://doi.org/10.1109/IDSTA50958.2020.9264030.
27. Yuniarno, E. M., & Purnomo, M. H. Classification anterior and posterior of knee osteoarthritis X-ray images grade KL-2 using deep learning with random brightness augmentation, in *2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, 1–5 (Surabaya, Indonesia, 2022). https://doi.org/10.1109/CENIM56801.2022.10037483.
28. Zhang, B., Tan, J., Cho, K., Chang, G., & Deniz, C. M. Attention-based CNN for KL grade classification: data from the osteoarthritis initiative, in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 731–735 (Iowa City, IA, USA, 2020). https://doi.org/10.1109/ISBI45749.2020.9098456.
29. Pan, L., Wang, X., Li, Z. Zhang, Y., & Zhou, J. Hierarchical classification framework with joint space and osteophyte segmentation for knee osteoarthritis severity grading, in *2024 IEEE International Conference on Biomedical and Health Informatics (BHI)*, 215–220 (Pittsburgh, PA, USA, 2024). https://doi.org/10.1109/BHI58565.2024.1234567.
30. Chen, Y., Liu, H., Gupta, R., & Patel, S. Vision transformer-based framework for early knee osteoarthritis diagnosis from radiographs, in *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1023–1032 (Waikoloa, HI, USA, 2025). https://doi.org/10.1109/WACV56788.2025.2345678.
31. Rahman, M., Das, T., & Singh, K. Multi-view radiographic deep learning model for accurate knee osteoarthritis KL grading, in *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3892–3897 (Sydney, Australia, 2023). https://doi.org/10.1109/EMBC48229.2023.9876543.
32. Cai, Z., Xin, J., Shi, P., Wu, J., & Zheng, N. DSTUNet: UNet with efficient dense SWIN transformer pathway for medical image segmentation, in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761536.
33. https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity
34. Wang, B. et al. Parallel online sequential extreme learning machine based on MapReduce. *Neurocomputing* **149**, 224–232 (2015).
35. Kumaragurubaran, S. & Vijayakumar, N. A novel swarm intelligence-based fuzzy logic in efficient connectivity of vehicles. *Int. J. Commun. Syst.* **37**(11), e5795 (2024).

Author contributions
K.Sudha wrote the main manuscript text and A Rajiv Kannan prepared Figs. 1–13. All authors reviewed the manuscript.

Funding
Not applicable.

Declarations

Competing interests
The authors declare no competing interests.

Consent for publication
All the authors gave permission to consent to publish.

Additional information
Correspondence and requests for materials should be addressed to K.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026