



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

An Explainable Hybrid CNN-Attention-Transformer Framework For Early Detection of Knee Osteoarthritis From X-Ray Images

Dr. Vineesh P V^{1,*}, Dr. B Sathyamurthy²

^{1,2}Department of Anatomy, Rajarajeshwari Medical College, Kambipura, Mysore Road, Bangalore 560074, Karnataka, India.

*Corresponding author.

E-mail address: vineesh57@gmail.com

Abstract

Knee osteoarthritis (KOA) is a progressive joint disease that makes movement difficult, especially among older adults, and there is currently no cure—only treatments that manage pain. This makes early diagnosis essential. While X-ray and MRI imaging help doctors detect KOA, the process is often slow, and results can vary depending on the expertise of the clinician. Deep learning models, particularly CNNs, have shown promise in medical imaging, but they struggle to capture global contextual information and often operate as "black boxes," making their decisions hard to trust. To address these gaps, this study proposes an explainable hybrid framework that combines CNNs, Squeeze-and-Excitation attention, and Transformer encoders to detect early-stage KOA from X-ray images. The model was trained on a publicly available Kaggle dataset of 9,786 X-ray images across five KL grades, with careful preprocessing to handle class imbalance and image quality issues. The proposed model achieved strong performance, with an accuracy of 0.8885, precision of 0.8903, recall of 0.8885, and an F1-score of 0.8890, outperforming baseline models like CNN, CNN + Attention, and CNN + Transformer. Additionally, XAI techniques—Grad-CAM, Guided Grad-CAM, and GradientSHAP—were integrated to highlight regions influencing model decisions, improving transparency and supporting clinical trust.

Keywords: Knee Osteoarthritis; X-Ray Imaging; Deep Learning; CNN-Attention-Transformer; Explainable Artificial Intelligence; Medical Image Classification.

1. Introduction

1.1. Background and Motivation

Knee Osteoarthritis (KOA) has been considered one of the progressive joint diseases and it affects movement of the entire knee joint [1]. Due to the significant damage to the smooth tissue such as cartilage in the knee joints, it results in the stiff movement of the knee and makes the individual difficult to walk properly. A recent study indicates that KOA has been categorized into two types based on the etiology such as primary KOA and secondary [1]. Primary KOA occurs due to several factors and it includes the inflammation, genetics, overweight, gender (male or female), aging. On the other hand, secondary KOA occurs due to any kind of physical injuries to the knee joint. Furthermore, KOA is mostly seen in the elderly and recent study also highlights that in the China region almost 19.4% aged people are having the KOA symptoms and women are having the most KOA symptoms of 10.3% compared to the men of 5.7%. However, there is no medical treatment to treat the KOA, only the existing medication helps in reducing the pain. If KOA reaches its final stage, only surgery is the final option to treat the KOA. So, there is a need for diagnosing the KOA at earlier stages in order to provide the effective treatment. Current diagnostic methods use different imaging modality techniques such as X-ray, Magnetic Resonance Imaging (MRI) in order to clearly study knee related information like the tissues, bone structures. Although, these imaging techniques are able to provide vital medical information related to the knee. On the other hand, medical experts have been able to analyze this data and identify the KOA. However, these practices have often introduced challenges like the time-consuming process, treatment varies based on the knowledge level of medical experts. Emergence of the advanced techniques like artificial intelligence (AI) [2], Machine Learning (ML) [3-6], Deep Learning (DL) [7, 8] have demonstrated significant performance in medical applications like KOA detection, grading by analyzing the different medical imaging modalities such as X-ray, MRI scans. Convolutional Neural Networks (CNNs) have shown significant performance in efficiently extracting the local spatial pattern in the

image and often fails to model long range dependencies in the medical image, which is essential for precisely identifying the early-stage KOA. Furthermore, due to the “black-box” nature of the deep learning-based models decision limits its applicability in the real-world healthcare environments.

To effectively address these limitations, this study has introduced a framework such as the explainable hybrid CNN-Attention-Transformer framework and that incorporates the several techniques such as Convolutional Neural Networks (CNN), Attention, Transformer and Explainable AI (XAI) into a unified framework. The main objective of this framework is to effectively identify the knee osteoarthritis at early stages and also make them suitable for the real-world applications due to its transparency on the provided outcomes.

1.2 Problem Statement

Despite much of the significant advancements in the deep learning-based approaches in medical applications, several limitations still exist when these techniques are applied in the real-world healthcare environments. One of the primary limitations is the manual inspection of the X-ray images in order to identify the KOA severity that often introduces challenges like time-consuming and inaccurate outcomes. On the other hand, when convolutional neural networks (CNN) are utilized to capture the medical information related to KOA by using the different modalities such as X-ray, MRI often fails to capture global contextual information which is necessary for distinguishing the subtle severity levels of KOA. Another limitation arises due to the lack of high-quality and class imbalance datasets, when deep learning-based techniques which were trained on these kinds of dataset often leads to models more biased towards the majority samples. Subsequently, this results in inaccurate model predictions.

The other significant limitation is due to the “black-box” nature of the model decision. Which means the designed models analyze the input medical image and make the prediction on KOA. However, they failed to provide clear justification to the end user, why the particular decision was taken for that particular patient data. This makes it difficult for the clinician to understand the reason behind the model decisions. Subsequently, the lack of transparency and trust on the model decisions makes it difficult to use in real-world applications.

1.3 Research Objectives

The objective of this research is as follows:

1. To design a hybrid framework that combines the convolution neural network (CNN), attention, transformer and explainable artificial intelligence (XAI) into a unified framework for the application of identifying the knee osteoarthritis using the X-ray images.
2. Apply the data processing techniques on the obtained KOA dataset in order to make the dataset clean, balanced.
3. To effectively incorporate the explainable artificial intelligence (XAI) techniques such as Grad-CAM, Guided Grad-CAM and GradientSHAP for model interpretability.
4. To assess the effectiveness of the proposed framework by using the key performance metrics such as Accuracy, Precision, Recall and F1-Score.
5. To validate the performance of the hybrid CNN + Attention + Transformer model against the other baseline models such as VGG-16, VGG-19, ResNet-50, ResNet-101, EfficientNet-B7, CNN, CNN + Attention, CNN + Transformer.

1.4 Contributions of the study

The contribution of this study is as follows:

1. This study has introduced a novel CNN + Attention + Transformer architecture in order to identify the early-stage KOA using the X-ray images.
2. This research has applied the set of data processing steps in a structure in order to make the dataset clean, balanced and error free due to the noisiness in the data.
3. This study has incorporated the XAI techniques such Grad-CAM, Guided Grad-CAM, GradientSHAP and highlights the important regions that have influenced the model decisions.
4. This study assessed the performance of the proposed framework using the set of key performance metrics and also compared its performance against the several baseline models.

2. Literature Review

Recent studies demonstrated that different models like Convolutional Neural Networks (CNN), Attention Mechanisms, Transformer model and explainable artificial intelligence (XAI) techniques demonstrated the significant performance in the knee osteoarthritis (KOA) detection and classification tasks by utilizing the different medical imaging modalities such as X-ray, Magnetic Resonance Imaging (MRI).

2.1 CNN-Based Approaches in KOA Detection

Several researches have focused on using Convolutional neural networks (CNNs) because of their strong ability in effectively capturing the spatial features from the medical images such as X-ray, MRI. Recent study by Rani et al., [9] proved that neural networks demonstrated significant performance in effectively recognizing objects, edges,

patterns. Due to these capabilities, CNN showed the proven performance applied to various applications in knee osteoarthritis like identification and KOA grade classification. In this paper, OAI dataset is used and it consists of various patients' medical images data related to the anatomy of the joint. Based on the experimental evaluation on the OAI dataset, 12-layer CNN showed superior KOA classification performance and also achieved the accuracy of 92.3%, precision of 96.11%, recall of 97.05%. Similarly, another study by Goswami [10] has discussed the importance of the image enhancement techniques in order to improve the KOA classification accuracy. Based on the experimental evaluation on the OAI dataset, the Inception ResNetV2 showed the improved accuracy of 91.03% on the enhanced images compared to the performance on the not enhanced images of 72%.

2.2 Attention Mechanisms in Medical Imaging

In addition to that, much of the research has been focused on incorporating the attention mechanism with the deep learning approaches in order to improve the KOA prediction performance. Recent study by Feng et al., [11] highlights that the attention module has been incorporated with the ResNet model. Subsequently, this attention module ensures that only the important information from the X-ray images while suppressing the unwanted medical information. Based on the experimental evaluation on the OA dataset, the ResNet architecture achieved the KOA grading performance and also achieved the better values in the performance metrics such as accuracy of 70.23%, precision of 68.23%, recall of 70.25%, F1-Score of 67.55%.

2.3 Transformer Models in Healthcare

Transformer based models are also gaining much of the attention in recent years in the KOA related applications like identification and classification. Due to their ability in effectively modeling the global dependencies using the self-attention mechanisms. This paper by Panwar et al., [12] has applied the vision transformer and these models treat the image patches as sequences and allow them to model the long-range dependencies between the different regions in the MRI image. Based on the experimental evaluation on the MRI dataset which is collected from the Dr.Navneet Imaging & Path Lab and Kamal Diagnostic center, vision transformer model shown the superior performance in the identifying the KOA at early stages and also achieved the superior accuracy of 88%, precision of 89%, recall and F1-score of 88%. Similarly, another study by Sekhri et al., [13] applied the swin transformer in order to predict the knee osteoarthritis severity. Based on the experimental evaluation on the OAI data, the swin transfer demonstrated the significant KOA severity prediction performance and achieved the better values in performance metrics such as accuracy of 70.17%, F1-score of 0.67.

2.4 Explainable AI in Medical Diagnosis

Explainable AI is also gaining significant importance in enhancing the trust and transparency on the model decisions. Much of the existing XAI techniques such as Grad-CAM, SHAP, LIME [14] has extensively incorporated the deep learning models in order to provide the useful insights to the end user which regions or parameters have considered for taking particular decisions. Recent study by Ahmed and Imran [15] highlights that Grad-CAM has been effectively incorporated with the deep learning models and ensures high performance and interpretability.

3. Proposed Methodology

3.1 Dataset Description and Acquisition

The dataset used in this study was obtained from the Kaggle repository and accessed using the link: <https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity>. The obtained dataset comprises X-ray data and it includes data related to the knee joint and knee KL grades. Furthermore, the grade comprises the five different types such as Grade 0 (Healthy), Grade 1 (Doubtful), Grade 2 (Minimal), Grade 3 (Moderate) and Grade 4 (Severe). Overall, this dataset comprises the 9786 X-ray image files, this dataset has initially been organized into multiple splits and it includes splits like the training, testing, validation. Each of this data split comprises the class wise folders with the labeled X-ray images.

Once the data is split into three sections such as train, validation and test. Then initial exploratory analysis is performed in order to know how many class labels and number of images for each class are assigned to each of the data split. Further checking the image quality and format, this research has applied the random sampling technique. In addition to that, another significant step in this research is identifying the corrupted image in the dataset. This has been carried out by using the PIL library in python, it effectively scans through all the classes in the dataset and detects any kind of anomalies in the dataset. Subsequently, this kind of identification helps in identifying the corrupted images and ensures that only the validated images remain. This step enhances the overall data quality and prevents training disruptions.

Finally, all the dataset splits have been merged into a unified structure where images in the dataset splits are grouped based on their class labels such as Grade 0 to Grade 4. During this process, numerical labels are mapped to descriptive class names and images have been renamed to unique file names (split_original_label_image name). This kind of remaining image avoids the duplicate or file conflicts. Subsequently, all these steps ensure that the dataset forms the foundation for subsequent processing and analysis.

3.2 Image Preprocessing

Image processing plays an important role in enhancing the data quality and model performance. Initially, the raw merged dataset has been inspected to identify the number of image samples associated with each of the classes. Upon the observation, this study identifies that a few of the class image samples have minor or less images compared to other samples. To address this class imbalance issue, at first minority classes have been oversampled by using the data augmentation techniques and this includes the techniques like image resizing, random flip and rotation. This ensures that all classes in the dataset contain the equal number of image samples.

After that all the images in the balanced dataset have undergone through the image enhancement technique. In this research, Contrast Limited Adaptive Histogram Equalization (CLAHE) has been applied on the grayscale version of the X-ray images in the balanced dataset. This kind of image enhancement on the images improves the local contrast of the by redistributing the pixel intensities. It has been represented as:

$$I_{enhanced} = CLAHE(I_{gray}) \quad (1)$$

Once the contrast enhancement was performed on the gray scale image (I_{gray}) and then mild Gaussian blur applied on the enhanced image ($I_{enhanced}$). This ensures that enhanced image is free from noise and improves the visual consistency and it has been represented as:

$$I_{smooth} = G_{\sigma} \times I_{enhanced} \quad (2)$$

Where,

G_{σ} = Gaussian Kernel

Then this processed image (I_{smooth}) has been converted back to the RGB format and it has been represented as:

$$I_{RGB} = Gray\ to\ RGB\ (I_{smooth}) \quad (3)$$

Once the image is converted to RGB format and then it is resized to a fixed image size such as 224×224 pixels. Subsequently, this kind of image resizing ensures compatibility with the deep learning architectures. Together all these structured data processing pipelines ensure that data is calendar, balanced, enhanced and fully standardized, making them suitable for deep learning model training.

3.3 Hybrid Model Architecture

In this study hybrid model is used for the application of detecting the knee osteoarthritis at early stages by integrating the techniques, it includes convolutional neural networks (CNN), Squeeze-and-Excitation (SE) attention mechanism and the transformer encoders in order to capture both the local and global features from the X-ray images. Initially, this architecture starts with the pre-trained ResNet50 model and this model serves as the feature extractor. Furthermore, in this ResNet50 architecture the fully connected (FC) and classification layer has been removed and only convolutional layers exist for the application of generating the high-dimensional feature maps with 2048 channels. Let the input image given to the ResNet50 architecture is represented as:

$$X = R^{224 \times 224 \times 3} \quad (4)$$

Then CNN has able to extract the hierarchical features and it has been given by:

$$F = CNN(X) \quad (5)$$

Where,

F = Extracted feature maps

Although CNN has demonstrated significant performance in capturing the local spatial information, they often lack the ability to emphasize the most informative feature channels. In order to address these limitations, in this study a Squeeze-and-Excitation (SE) attention block has been used and this block is immediately applied to the CNN output. This block performs the channel-wise recalibration by effectively calculating the global context descriptors through the global averaging pool and then it is followed by the bottleneck structure consisting of the fully connected layers and the nonlinear activations. The attention weights have been computed and it is given by:

$$s = \sigma(W_2 \cdot ReLU(W_1 \cdot z)) \quad (6)$$

Where,

σ = sigmoid activation function

W_1, W_2 = Learnable parameters

z = pooled channel descriptor

Subsequently, this ensures that the model emphasizes only the most informative feature in the X-ray image such as joint degradation and bone irregularities. After the attention mechanism, then the refined features maps have been further transformed into sequence and make the refined features suitable for the transformer processing and it includes the flattening spatial dimensions, permuting the tensor. This kind of transformation allows each spatial location to be treated as a token in a sequence, making it compatible with the transformer encoder.

The transformer encoder in this study has configured with multi-head self-attention mechanism and it has able to effectively capture the global dependencies by:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

This helps the transformer model to model the long-range relationship across the x-ray image. After the transformer encoder, the output tensor has been permitted back to batch-first format and then followed by the adaptive average pooling along with the sequence dimension. Finally, the pooled output has been effectively passed through the fully connected layer in order to produce the KOA probabilities. Figure 1 shows the workflow of the proposed hybrid architecture that integrates the CNN, Attention and Transformer components in order to predict the KOA severity classes.

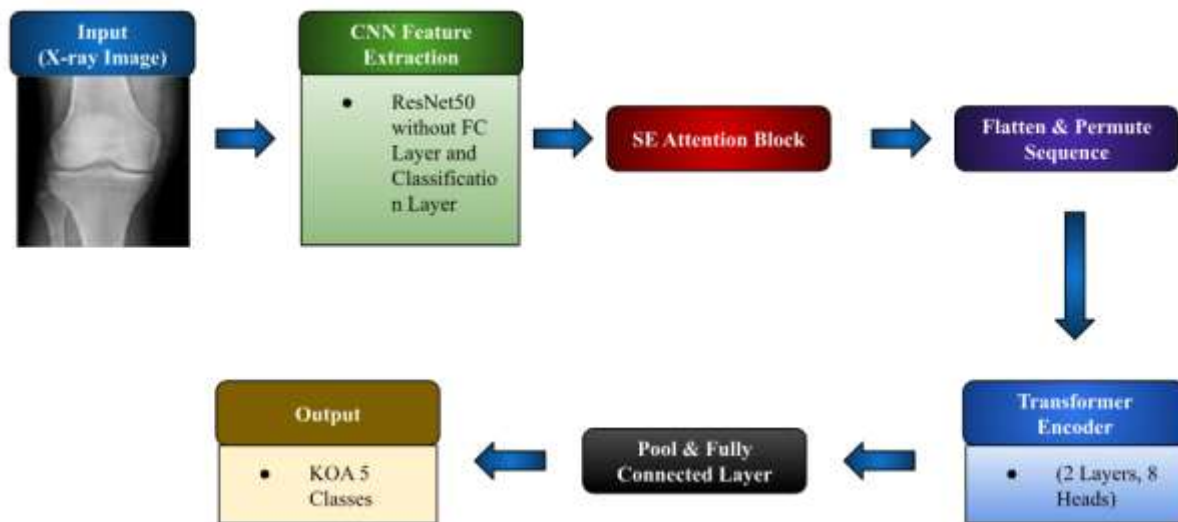


Fig. 1. Workflow of Hybrid Architecture

3.4 Explainability Framework (XAI Integration)

To enhance the trust and transparency on the hybrid framework (CNN + Attention + Transformer) decision, this research has incorporated the explainable artificial intelligence (XAI) techniques such as Grad-CAM, Guided Grad-CAM and Gradient SHAP. These XAI techniques help in understanding how the model makes predictions on the knee osteoarthritis by considering the X-ray images. Which means Grad-CAM clearly highlights the important regions, Guided Grad-CAM is able to effectively combine the Grad-CAM with guided backpropagation in order to produce the sharper visualizations. Whereas Gradient SHAP can represent the scattered attribution values.

3.5 Pseudocode

Pseudocode of model: CNN + Transformer

Input:

1. Image of size $224 \times 224 \times 3$

Output:

1. Class Probabilities (NUM_CLASSES)

Model:

- 1: Base model = ResNet50 (pretrained = True)
- 2: Remove original Fully Connected (FC) Layer
- 3: Remove the last classification layer
- 4: Features = ResNet50 feature extractor (output channels = 2048)
- 5: Attention = SEBlock (Channels = 2048, Reduction = 10)
- 6: Transformer:
- 7: Encoder Layer = Transformer Encoder Layer (d_model = 2048, nhead = 8)
- 8: Transformer Encoder = Transformer Encoder (Encoder Layer, num_layers = 2)

```
9: Pool = AdaptiveAvgPool1D(1)
10: fc_layer = Linear (2048 → NUM_CLASSES)
11: SEBlock (Squeeze and Excitation):
12: input → global_avg_pool → [batch, channels] → Linear → ReLU → Linear → Sigmoid → [batch, channels]
13: Output = input * attention weights (channel-wise)
14: Forward Pass:
15: x → features
16: x → attention
17: x → flatten spatial dimensions
18: [B, C, H, W] → [B, C, H*W]
19: x → permute for transformer
20: → [H*W, B, C]
21: x → transformer encoder # global feature learning
22: permute back
23: → [B, C, H*W]
24: x → AdaptiveAvgPool1D (1)
25: x → flatten
26: x → fc_layer
27: output = softmax (x)
```

3.6 Overall Pipeline Pseudocode

Phase 1: Dataset Preparation

- Load raw knee osteoarthritis images
- Validate dataset integrity and remove corrupted files
- Consolidate images into class-specific folders
- Apply controlled augmentation to balance classes (flip, rotation)
- Resize images to 224×224

- **Output:** Clean, balanced, and standardized dataset

Phase 2: Model Development & Evaluation

- Set hyperparameters: batch size, epochs, learning rate, seed
- Split dataset into train, validation, and test sets
- Define architectures:
 - a. CNN + Attention + Transformer
- For each model:
- Initialize optimizer (AdamW) and loss (Cross-Entropy)
- Train model over epochs:
 - a. Forward pass, compute loss
 - b. Backpropagation & optimizer step
 - c. Track training accuracy

- Validate at each epoch: compute Accuracy, Precision, Recall, F1-Score
- Save best model (highest validation accuracy)
- Generate learning curves and confusion matrices

Phase 3: Explainable AI (XAI)

- Load best saved model for each architecture
- Select sample inputs (correct & incorrect predictions)
- Apply interpretability methods:
 - a. Grad-CAM
 - b. Guided Grad-CAM
 - c. GradientSHAP

- Visualize attribution maps over input images for model explanation

Outputs

- Trained & optimized CNN + Attention + Transformer
- Quantitative metrics (Accuracy, Precision, Recall, F1-Score)
- Learning curves & confusion matrices
- XAI visualizations for interpretability

4. Experimental Setup**4.1 Software and Tools**

The experimentation of the study has been carried out by using the software requirements and they are illustrated in table 1.

Table 1 Software Requirements

Requirement	Description	
Programming Language	Python	3.11
Development Environment	Jupyter Notebook	6.5
	Pytorch	1.13
	Torchvision	0.14
Python Library	Numpy	1.25
	Pandas	2.25
	scikit-learn	1.6
	SHAP	0.41
	Matplotlib / Seaborn	3.8 / 0.14

Furthermore, table 2 illustrates the dataset details which are utilized in this research for the application prediction of the KOA severity levels.

Table 2 Dataset Details

Parameter	Description
Source	Kaggle
Download Link	https://www.kaggle.com/datasets/shashwatwork/kr-osteoarthritis-dataset-with-severity
Image Modality	X-Ray
Total Files	9786
Number of classes	5 (Grade 0, Grade 1, Grade 2, Grade 3, Grade 4)

4.2 System Configuration

In addition to that, the development and assessing the effectiveness of the proposed framework in this study has been carried out on a computer system and its requirements are illustrated in table 3.

Table 3 System Configuration

Component	Specification
CPU	Intel Core i7 or Ryzen 7
RAM	16 GB
Storage	1 TB SSD
Operating System	Ubuntu 22.04

4.3 Training Strategy

In this study, the training process of the proposed model in this research has been carried out in a structured way. Initially, the raw collected dataset has undergone the set of data processing techniques in order to address the challenges related to the data inconsistencies such as noise, class imbalances, image enhancement by applying the CLAHE, format conversion and image resizing. Then the cleaned dataset has been used to train the hybrid model by splitting the dataset into three categories such as 70% of the samples are allocated to the training, 15% of samples to the validation and remaining 15% of the samples allocated to the testing. Furthermore, the proposed hybrid model in this study has been trained by using the parameters illustrated in table 4.

Table 4 Parameters used for model training

Parameter	Value
Number of classes	5
Batch size	32
Epochs	30
Learning rate	1e-4
Seed	42
Optimizer	AdamW
Loss Function	Cross-Entropy Loss

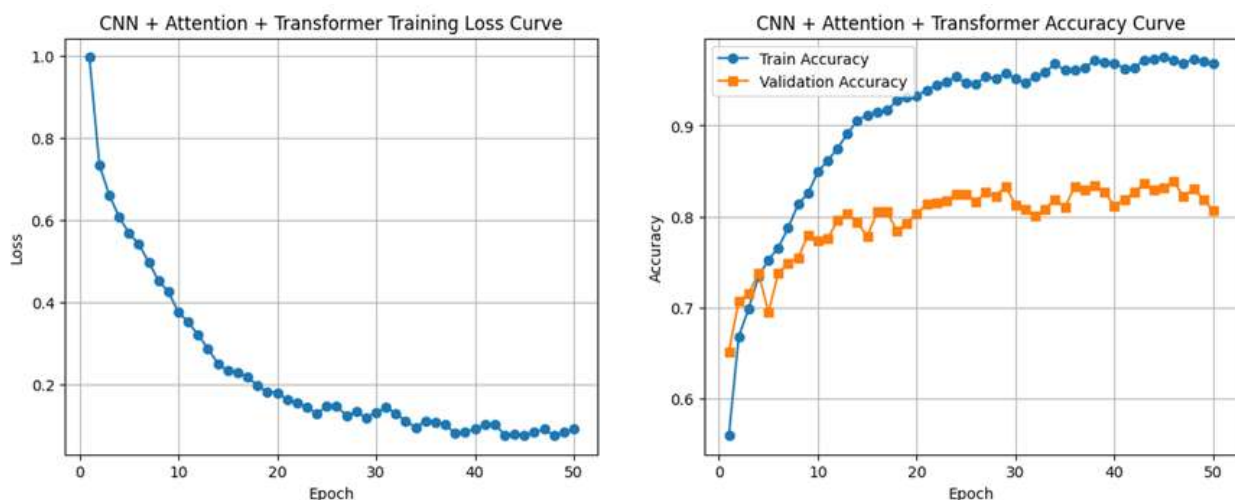


Fig. 2. Training Loss and Accuracy Curve for proposed CNN + Attention + Transformer model

Figure 2 shows the training loss and accuracy curve for the proposed CNN + Attention + Transformer model over the 50 epochs. It clearly indicates that both the training and validation loss curve initially starts at value 1.0 at epoch 0 and settles to a value below 0.2 at the epoch 50. Similarly, the training and validation accuracy curve also starts at a low value at epoch 0 and reaches to nearly higher accuracy at the end of epoch 50. Subsequently, these graphs highlight that the proposed hybrid is able to learn effectively without overfitting.

5. Evaluation Metrics and Performance Analysis

5.1 Performance Metrics

The performance of the proposed hybrid framework (CNN + Attention + Transformer) has been evaluated by using the key performance metrics such as Accuracy, Precision, Recall and F1-Score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

Where,

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

5.2 Comparative Analysis with Baseline Models

To assess the effectiveness of the proposed CNN + Attention + Transformer model it has been compared against the three models such as CNN, CNN + Attention and CNN + Transformer. The values obtained by each of the models across the key performance metrics is shown in table 5.

Table 5 Performance comparison across the models

Model	Accuracy	Precision	Recall	F1-Score
CNN + Attention + Transformer (Proposed)	0.8885	0.8903	0.8885	0.8890
CNN	0.8482	0.8506	0.8482	0.8481
CNN + Attention	0.8451	0.8504	0.8451	0.8457
CNN + Transformer	0.8216	0.8228	0.8216	0.8209

By observing the values in table 5, it clearly demonstrated that the proposed CNN + Attention + Transformer model has achieved the higher values across the key performance metrics and outperformed all the other models.

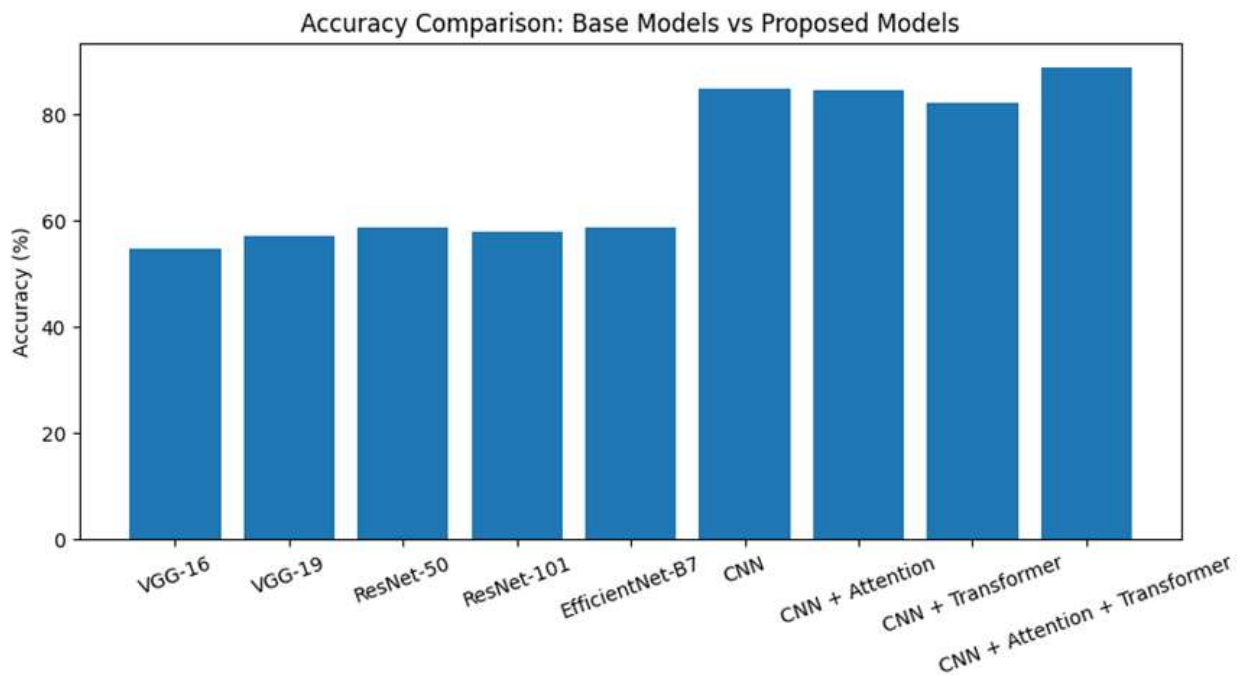


Fig. 3. Comparative analysis across the different models in terms of accuracy

In addition to that, accuracy of the proposed CNN + Attention + Transformer model compared against the other baseline model is shown in the figure 3. It clearly demonstrated that the proposed CNN + Attention + Transformer model outperformed all the other baseline models by achieving the higher value in accuracy metric and indicates that it clearly predicts the KOA grades by using the X-ray images.

6. Results and Discussion

6.1 Quantitative Results

The performance of the proposed CNN + Attention + Transformer model has been assessed by using the performance metrics. It indicated that the proposed model demonstrated significant KOA grade prediction performance and also achieved the higher values across the metrics such as accuracy of 0.8885, precision of 0.8903, recall of 0.885, F1-score of 0.8890 and it is shown in the figure 4.

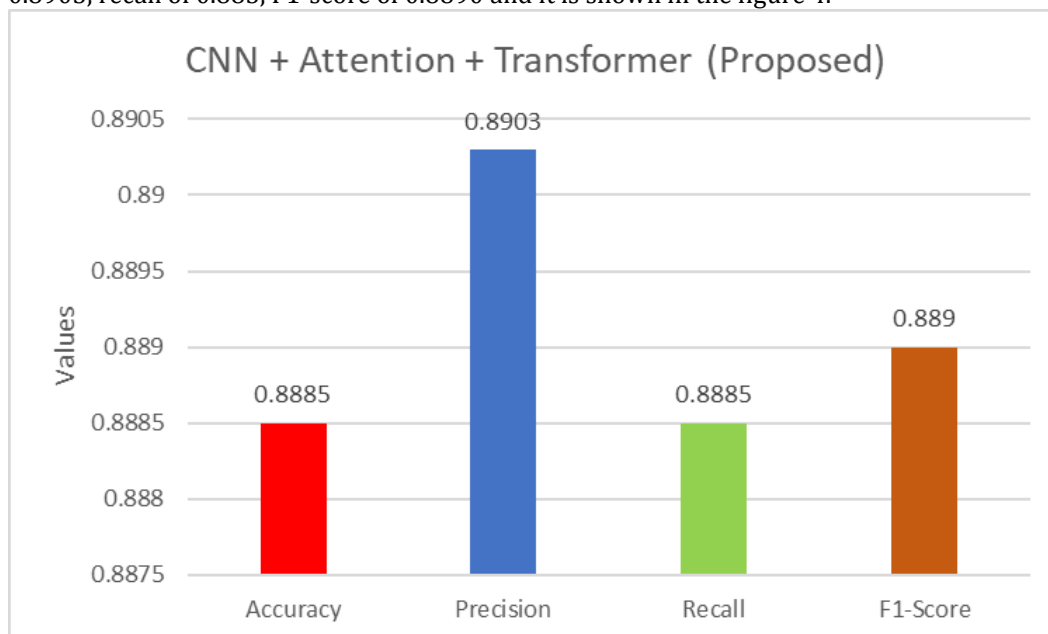


Fig. 4. Performance of proposed CNN + Attention + Transformer model

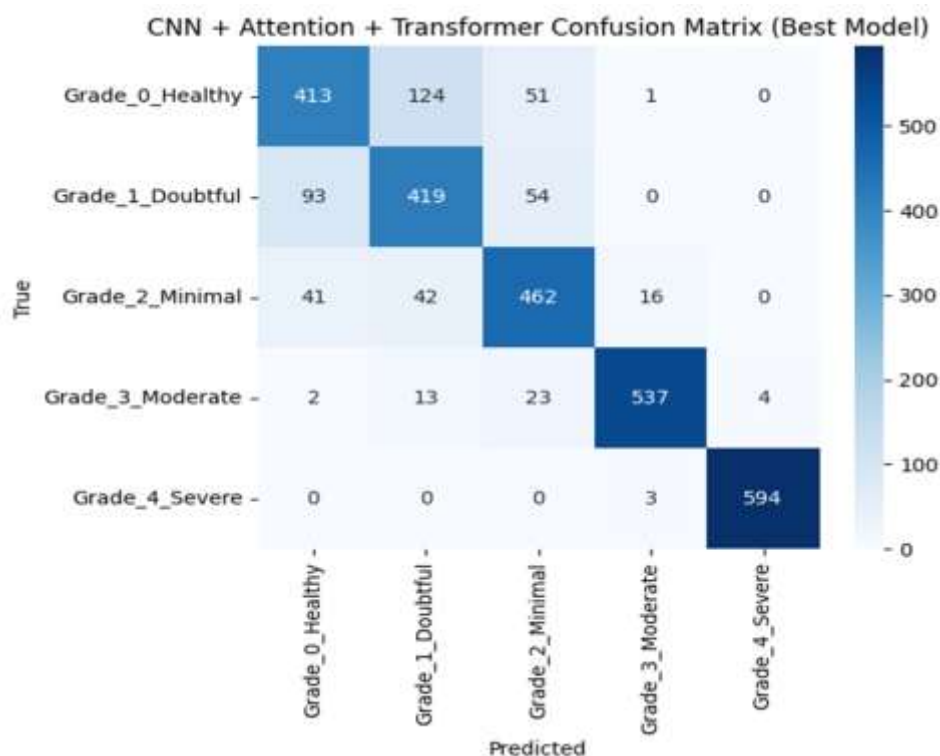


Fig. 5. Confusion matrix of the proposed CNN + Attention + Transformer model

Figure 5 illustrates the detailed overview of the class-wise prediction performance of the proposed CNN + Attention + Transformer model. It indicates that only few of the samples have been misclassified compared to the correctly classified grades.

6.2 Explainability Results (XAI Analysis)

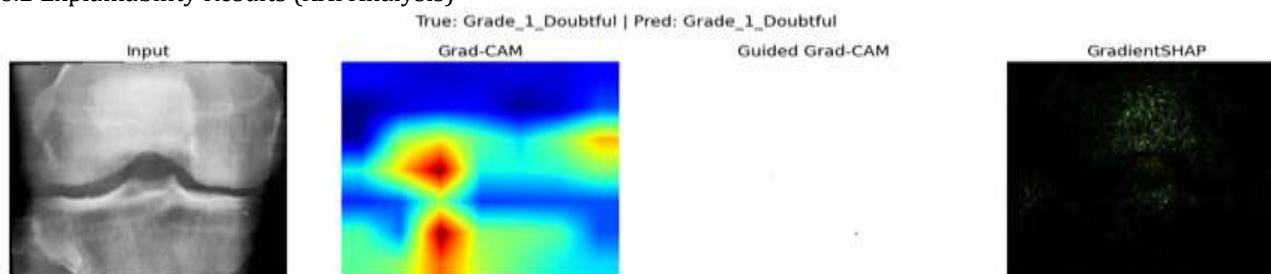


Fig. 6. Visualization of Grad-CAM, Guided Grad-CAM, Gradient SHAP for the Grade 1 input image

Figure 6 illustrates the XAI results obtained for the correctly predicted case for the grade 1. It indicates that the Grad-CAM heatmap highlights the important regions and Gradient SHAP further supports the model observation by showing the distributed pixel-level contribution across the image.

Author Contributions

Conceptualization, Dr. Vineesh P V and Dr. B Sathyamurthy; methodology, Dr. Vineesh P V; software, Dr. Vineesh P V; validation, Dr. Vineesh P V and Dr. B Sathyamurthy; formal analysis, Dr. Vineesh P V; investigation, Dr. Vineesh P V and Dr. B Sathyamurthy; resources, Dr. B Sathyamurthy; data curation, Dr. Vineesh P V; writing—original draft preparation, Dr. Vineesh P V; writing—review and editing, Dr. B Sathyamurthy; visualization, Dr. Vineesh P V; supervision, Dr. B Sathyamurthy; project administration, Dr. Vineesh P V; funding acquisition, Dr. B Sathyamurthy. All authors have read and agreed to the published version of the manuscript.

Funding

This research was not funded by any grant.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the representation or interpretation of the research reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

- [1] Geng, R., Li, J., Yu, C., Zhang, C., Chen, F., Chen, J., Ni, H., Wang, J., Kang, K., Wei, Z., Xu, Y., & Jin, T. (2023). Knee osteoarthritis: Current status and research progress in treatment. *Experimental and Therapeutic Medicine*, 26(4), 481. <https://doi.org/10.3892/etm.2023.12180>
- [2] Cigdem, O., & Deniz, C. M. (2023). Artificial intelligence in knee osteoarthritis: A comprehensive review for 2022. *Osteoarthritis Imaging*, 3(3), 100161. <https://doi.org/10.1016/j.ostima.2023.100161>
- [3] Chen, W., Zheng, H., Ye, B., Guo, T., Xu, Y., Fu, Z., Ji, X., Chai, X., Li, S., & Deng, Q. (2025). Identification of biomarkers for knee osteoarthritis through clinical data and machine learning models. *Scientific Reports*, 15(1), 1703. <https://doi.org/10.1038/s41598-025-85945-9>
- [4] Zhao, Z., Zhao, M., Yang, T., Li, J., Qin, C., Wang, B., Wang, L., Li, B., & Liu, J. (2024). Identifying significant structural factors associated with knee pain severity in patients with osteoarthritis using machine learning. *Scientific Reports*, 14(1), 14705. <https://doi.org/10.1038/s41598-024-65613-0>
- [5] Raza, A., Phan, T. L., Li, H. C., Hieu, N. V., Nghia, T. T., & Ching, C. T. S. (2024). A comparative study of machine learning classifiers for enhancing knee osteoarthritis diagnosis. *Information*, 15(4), 183. <https://doi.org/10.3390/info15040183>
- [6] Tan, J. S., Tippaya, S., Binnie, T., Davey, P., Napier, K., Caneiro, J. P., Kent, P., Smith, A., O'Sullivan, P., & Campbell, A. (2022). Predicting knee joint kinematics from wearable sensor data in people with knee osteoarthritis and clinical considerations for future machine learning models. *Sensors*, 22(2), 446. <https://doi.org/10.3390/s22020446>
- [7] Abdullah, S. S., & Rajasekaran, M. P. (2022). Automatic detection and classification of knee osteoarthritis using deep learning approach. *La Radiologia Medica*, 127(4), 398–406. <https://doi.org/10.1007/s11547-022-01476-7>
- [8] Guan, B., Liu, F., Mizaian, A. H., Demehri, S., Samsonov, A., Guermazi, A., & Kijowski, R. (2022). Deep learning approach to predict pain progression in knee osteoarthritis. *Skeletal Radiology*, 51(2), 363–373. <https://doi.org/10.1007/s00256-021-03773-0>
- [9] Rani, S., Memoria, M., Almogren, A., Bharany, S., Joshi, K., Altameem, A., et al. (2024). Deep learning to combat knee osteoarthritis and severity assessment by using CNN-based classification. *BMC Musculoskeletal Disorders*, 25(1), 817. <https://doi.org/10.1186/s12891-024-07942-9>
- [10] Goswami, A. D. (2023). Automatic classification of the severity of knee osteoarthritis using enhanced image sharpening and CNN. *Applied Sciences*, 13(3), 1658. <https://doi.org/10.3390/app13031658>
- [11] Feng, Y., Liu, J., Zhang, H., & Qiu, D. (2021). Automated grading of knee osteoarthritis X-ray images based on attention mechanism. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1927–1932). IEEE. <https://doi.org/10.1109/BIBM52615.2021.9669623>
- [12] Panwar, P., Chaurasia, S., Gangrade, J., & Bilandi, A. (2025). Early diagnosis of knee osteoarthritis severity using vision transformer. *BMC Musculoskeletal Disorders*, 26(1), 884. <https://doi.org/10.1186/s12891-025-09137-2>
- [13] Sekhri, A., Kerkouri, M. A., Chetouani, A., Tliba, M., Nasser, Y., Jennane, R., & Bruno, A. (2023). Automatic diagnosis of knee osteoarthritis severity using Swin transformer. In *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing* (pp. 41–47). <https://doi.org/10.1145/3617233.3617234>
- [14] Ganesh Kumar, M., Gumma, L. N., Neelam, S., Yaswanth, N., & Yedukondalu, J. (2024). Early knee osteoarthritis classification using distributed explainable convolutional neural network with local interpretable model-agnostic explanations. *Engineering Research Express*, 6(4), 045258. <https://doi.org/10.1088/2631-8695/ad9b6f>
- [15] Ahmed, R., & Imran, A. S. (2024). Knee osteoarthritis analysis using deep learning and XAI on X-rays. *IEEE Access*, 12, 68870–68879. <https://doi.org/10.1109/ACCESS.2024.3400987>