IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 July 2024, accepted 24 July 2024, date of publication 29 July 2024, date of current version 14 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435572

RESEARCH ARTICLE

KOA-CCTNet: An Enhanced Knee Osteoarthritis Grade Assessment Framework Using Modified Compact Convolutional Transformer Model

MUSHRAT JAHAN[1], MD. ZAHID HASAN[1], (Member, IEEE), ISMOT JAHAN SAMIA[1], KANIZ FATEMA[1], MD. AWLAD HOSSEN RONY[1], MOHAMMAD SHAMSUL AREFIN[2], (Senior Member, IEEE), AND AHMED MOUSTAFA[3,4,5]

[1]Health Informatics Research Laboratory, Department of Computer Science and Engineering, Daffodil International University, Dhaka 1341, Bangladesh
[2]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh
[3]School of Psychology, Faculty of Society and Design, Bond University, Gold Coast, QLD 4226, Australia
[4]Department of Human Anatomy and Physiology, Faculty of Health Sciences, University of Johannesburg, Johannesburg 2092, South Africa
[5]Centre for Data Analytics, Bond University, Gold Coast, QLD 4226, Australia

Corresponding author: Md. Zahid Hasan (zahid.cse@diu.edu.bd)

ABSTRACT Knee osteoarthritis (KOA) is a prevalent condition characterized by gradual progression, resulting in observable bone alterations in X-ray images. X-rays are the preferred diagnostic tool for their ease of use and cost-effectiveness. Physicians use the Kellgren and Lawrence (KL) grading system to understand the severity of an individual condition of KOA. This system categorizes the disease from normal to a severe stage. Early detection of the condition with this approach enables knee deterioration to be slowed down with therapy. In this study, we aggregated four datasets to generate an extensive dataset comprising 110,232 raw images by applying an augmentation technique called deep convolutional generative adversarial network (DCGAN). We employed advanced image pre-processing methods (adaptive histogram equalization (AHE), fast non-local means), including image resizing, to generate a substantial dataset and enhance image quality. Our proposed approach involved developing a modified compact convolutional transformer (CCT) model known as KOA-CCTNet as the foundational model. We further investigated optimal configurations by adjusting various parameters and hyperparameters in the final model to handle large datasets and address training time concerns efficiently. We investigated optimizing its configurations by adjusting numerous parameters and hyperparameters to efficiently manage extensive data and address concerns related to training time. Simulation results indicated that our proposed model outperforms other transfer learning models (Swin Transformer, Vision Transformer, Involutional Neural Network) in terms of accuracy. The test accuracy for the ResNet50, MobileNetv2, DenseNet201, InceptionV3, and VGG16 was 80.77%, 79.98%, 80.23%, 76.89%, and 79.58%, respectively. All of them were surpassed by our proposed KOA-CCTNet model, which had a test accuracy of 94.58% while classifying KOA X-ray images. Furthermore, we reduced the number of images to assess the model's performance and compared it to existing models. However, by employing a large datahub, our proposed approach provides a unique and effective way to diagnose KOA grades with satisfying results.

INDEX TERMS Knee osteoarthritis, deep convolutional generative adversarial network (DCGAN), knee osteoarthritis grades, image pre-processing, compact convolutional transformer, knee X-ray.

The associate editor coordinating the review of this manuscript and approving it for publication was You Yang.

I. INTRODUCTION

Knee osteoarthritis (KOA) is a common and disabling joint disease that has become twice as prevalent since the mid-20th century [1]. It occurs when the knee joint becomes

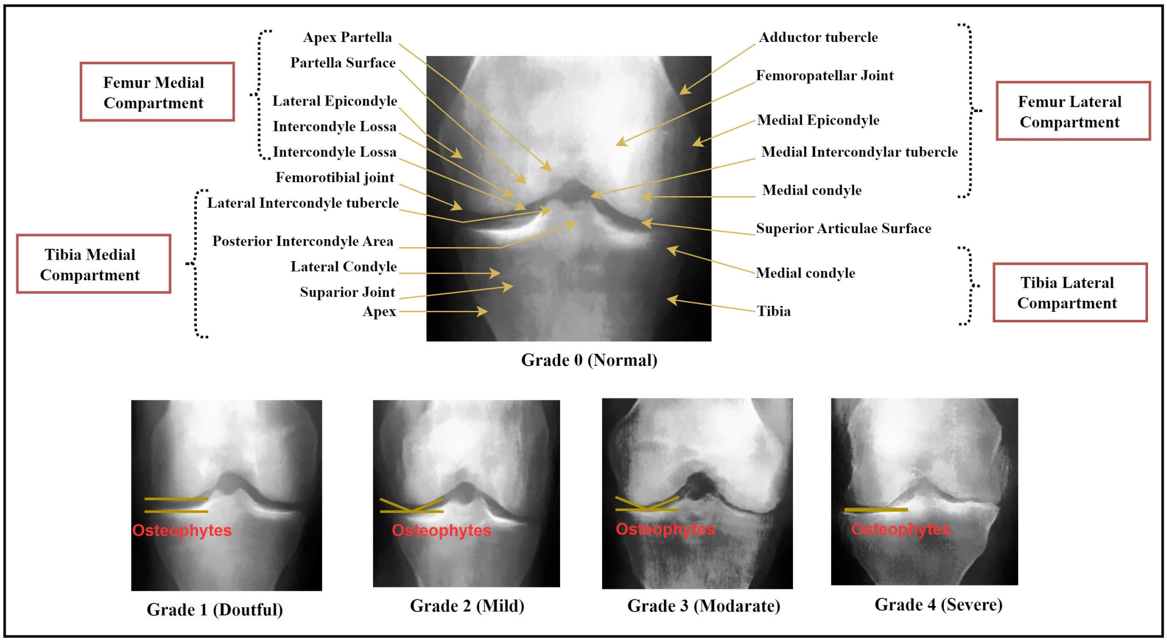


FIGURE 1. The example of the Kellgren–Lawrence (KL) scale.

inflamed due to regular wear and tear, causing the cartilage to deteriorate and become damaged [2]. It primarily targets the cartilage, the protective tissue covering the bone ends. Cartilage allows smooth joint movement in a healthy state, preventing bone-on-bone friction [3]. KOA involves the breakdown of the outer cartilage layer, resulting in painful bone friction [4]. This condition often impacts joints like the knee, spine, hip, and foot. There are two forms: primary and secondary. Primary KOA commonly affects older individuals and can be attributed to genetics or aging. On the other hand, secondary KOA is linked to factors such as injuries, diabetes, vigorous sports activities, and rheumatoid arthritis, and it typically manifests earlier in life. Based on the Global Burden of Disease (GBD) Study, the global prevalence of symptomatic osteoarthritis is 9.6% in men and 18% in women among individuals over the age of 60. Additionally, the study found that the prevalence of osteoarthritis is 16% in individuals aged 15 years and older, 22.99% in individuals aged 40 years and older [5]. According to separate research, the proportion of people with KOA in the 45+ age group is expected to rise from 13.8% to 15.7% by 2032 [6]. The primary symptoms of KOA include severe pain, limited joint mobility, and stiffness in the morning, which can significantly impact a patient's daily activities [7]. Diagnosing KOA typically involves assessing pain and considering a combination of clinical and radiographic symptoms [8]. However, measuring pain precisely is difficult since it is subjective. There are several methods for classifying the phases of osteoarthritis [9].

Furthermore, X-rays are readily accessible, cost-effective, and safe imaging tools that aid in identifying articular cartilage degeneration, narrowing of the space between bones, and the formation of bone spurs. The KL scale is a widely recognized radiographic classification system designed to gauge the severity of KOA knee osteoarthritis using X-rays. Depending On the severity of the symptoms of osteophytes and Joint Space Narrowing (JSN), KOA can be categorized into five grades. Grading begins at Grade 0, denoting a knee with no osteoarthritic narrowing and potential osteophytic lipping within this scale. Grade 2 confirms the existence of tiny osteophytes that may cause a reduction in joint space. Grade 3 showcases evident joint space reduction, multiple moderate osteophytes, some bone hardening, and potential deformities in bone contour [10]. Finally, Grade 4 is characterized by pronounced joint space narrowing, large osteophytes, intense bone sclerosis, and distinct bone deformities. In this regard, there is a demand for computer-assisted diagnostic tools to assist healthcare professionals in consistently and automatically assessing the severity of KOA. Since a patient's clinical symptoms might not always align perfectly with their KL grades, introducing variability in symptom experiences relative to radiographic findings is also crucial. That's why, before proposing any automated system, it's essential to understand the medical characteristics of the diseases. It will enhance the reliability and reproducibility of X-ray interpretation for KOA diagnosis by offering significant advantages [11]. However, Figure 1 and Table 1 show examples of using the KL scale to assess KOA severity and descriptions of the KOA grades.

However, the adoption of automation systems in the assessment of KOA is driven by their ability to enhance efficiency, accuracy, and accessibility in clinical settings [12]. Automated tools, particularly those incorporating artificial intelligence and machine learning, provide rapid and consistent analysis of medical images, significantly reducing the time required for diagnosis and ensuring uniformity in the evaluation of different patients [12], [13]. Additionally,

TABLE 1. Knee osteoarthritis (KOA) grade description.

Garde	Intensity	Statement
Grade 0	Normal	No radiographic features of KOA are present
Grade 1	Doubtful	Doubtful joint space narrowing and possible presence of small osteophytes.
Grade 2	Mild	Definite joint space narrowing and the presence of osteophytes
Grade 3	Moderate	Moderate joint space narrowing, multiple osteophytes, and possible bone remodeling
Grade 4	Severe	Severe joint space narrowing, extensive osteophyte formation, and possible joint deformity

a viable method for diagnosing KOA disorders might be an automated and trustworthy method based on deep learning algorithms employing X-ray images. Identifying and classifying KOA diseases using CNN models trained on knee x-ray images [22] requires a lot of resources and computing time. These obstacles are exacerbated by the issue of small medical datasets with an imbalanced number of images in the various classifications. However, convolutions are not necessary to create effective classification models that solve the computational complexity problem. In this sense, machine learning (ML) research has placed a significant emphasis on transformers. Vision Transformer (ViT), which applies a pure self-attention-based model to sequences of image patches and achieves competitive performance when compared to CNNs, is the most noteworthy development in this field [14]. In terms of computing efficiency and accuracy, ViT models beat CNNs by nearly a factor of four, and they achieve superior accuracy on large datasets with less training time. When we look at how hard it is to program the "sequential operations," transformers are much better than recurrent neural networks (RNNs) because self-attention layers [15] work faster than recurrent layers. ViTs can solve the problem of training time, but because of the transformer models' architecture, they are data-hungry and need a ton of data to function well. Gathering a significant volume of annotated image data for medical research is frequently difficult, expensive, and time-consuming. In order to address this, Hassani et al. [16] added basic convolutional blocks to the vision transformer's tokenization stage, introducing the Compact Convolutional Transformer (CCT). This resulted in a reduction of training time and a noticeable improvement in performance. This work utilizes four distinct datasets to automatically detect and classify KOA disorders into the Normal, Doubtful, Mild, Moderate, and Severe categories. This context admirably handles small medical datasets, an imbalanced number of images, low-resolution images, training duration and complexity, among other difficulties.

The following are the primary contributions of this research:

1. Differences in the quality and features of the images arise from gathering different X-ray image datasets using various sources and protocols. We establish a large data hub of 11,431 X-ray images by combining four datasets (Mendeley I [25], Mendeley II [26], Kaggle [27], and AIDA Data Hub [28]) with varying X-ray image attributes and sizes. This study addresses the challenge of handling the variation of the dataset.
2. The experiment's dataset exhibits varying class imbalances, which decrease the effectiveness of the model. To address this issue, we proposed a state-of-the-art deep convolutional generative adversarial network (DCGAN) augmentation method in this study.
3. Advanced image pre-processing techniques such as Adaptive Histogram Equalization (AHE) and Fast Non-Local Means (FNLM) are utilized to augmented images to eliminate noises and artifacts, enhance image quality, and address illumination variations. These techniques improve feature visibility, clarity and standardized dimensions of X-ray images.
4. To evaluate the model performance in handling large-scale image datasets, training speed, and accuracy, we propose a compact convolutional transformer (CCT) model named KOA-CCTNet. The model demonstrates remarkable classification performance and operating efficiency with large datasets and compare the propose model with other state of the art.

This paper is structured as follows: Section II presents a comprehensive review of existing literature. Then, a detailed discussion of the research methodology is provided in Section III, followed by Section IV, where we elucidate the results and analyse our proposed model utilizing a range of performance metrics. This section also includes a comparative analysis of existing transfer learning models and state-of-the-art works and an examination of our model's robustness. The paper is concluded in Section V, where we summarize our findings, and Section VI offers insights into the limitations of our study and potential directions for future research.

II. LITERATURE REVIEW

Many studies have reviewed the use of knee X-ray images to diagnose KOA. They have highlighted advancements in deep learning and machine learning for accurate diagnosis. In this section, we reviewed how these images have been used with different techniques to identify KOA.

Ganesh Kumar and Das Goswami [17] proposed a method for automatic classification of the severity of KOA using enhanced image sharpening and convolutional neural networks (CNN). The KL grading system is used to assess the severity of the KOA grades. The study used baseline X-ray images from the Osteoarthritis Initiative (OAI). Their proposed method achieved a mean accuracy of 91.03%, which is an improvement of 19.03% over the earlier accuracy of 72% by using the original knee X-ray images for the detection of OA with five gradings. Limitations of the study include the fact that the images taken are not pre-processed correctly, which impacts the model's overall performance. Additionally, The model's accuracy is not promising due to patients' objectivity, such as pain grade. Besides, Mohammed and his colleagues [18] discussed using deep neural network (DNN) models to detect and classify KOA using X-ray images. The authors used six pre-trained DNN models in their experiments: VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNetV2, and DenseNet121 and achieved maximum classification accuracies of 69%, 83%, and 89% on three different datasets. The ResNet101 DNN model outperformed among them.

In another study, researcher [19] introduced a deep learning computer technique to automatically spot and classify knee arthritis from X-ray images. Their method was nearly 99% accurate in identifying a specific area in the knee and determining the severity of the arthritis. They utilized technical models such as Faster RCNN to pinpoint this area and ResNet-50 to gather image features. They also used AlexNet to determine how bad the arthritis was. This new method was better than others, especially in identifying the early stages of arthritis.

In addition, El-Ghany et al. [20] presented a proposed model that showcased a remarkable 95.93% accuracy for multi-classification and 93.78% for binary classification in diagnosing KOA. The DenseNet169 model surpassed other deep learning techniques such as InceptionV3, Xception, ResNet50, and others in various performance metrics. Notably, while past research capped an accuracy of 87%, this model excelled in multi-classification and binary classification. Tested using the pre-processed OAI dataset and benchmarked against recent classifiers, the model's prowess lies in its capability to diagnose KOA severity from X-ray images precisely. The mentioned heightened accuracy paves the way for more effective and economical KOA diagnoses, ultimately leading to enhanced patient care and controlled disease progression. Another researcher [21] introduces an automated deep-learning technique to detect and classify KOA from X-ray images. Utilizing a dataset from the Osteoarthritis Initiative (OAI), it was divided into training, testing, and validation portions. By leveraging transfer learning, the authors refined models like ResNet-34, VGG-19, and DenseNet variants and combined them for better results. This approach achieved an impressive 98% accuracy and a Quadratic Weighted Kappa score of 0.99, especially improving the accuracy for specific arthritis grades. For context, the paper also references past studies that employed various techniques, positioning their results alongside this new method for a comprehensive comparison.

Nasser et al. [13] present the Discriminative Shape-Texture Convolutional Neural Network (DST-CNN) as a solution for early KOA detection using X-ray images. This model enhances classification using a unique discriminative loss and a Gram Matrix Descriptor (GMD) block to analyze texture and shape features. Remarkably, DST-CNN showcases top-tier results, boasting an accuracy of 74.08%, precision of 68.46%, and other robust metrics. It stands out significantly when differentiating between healthy individuals and borderline cases, outclassing other CNN-based methods. The paper underscores that each aspect of DST-CNN amplifies its efficiency, with the best outcomes when blending texture insights at various stages with the discriminative loss. Given the subtle differences in early-stage KOA X-ray images, this model emphasizes improved texture analysis via the GMD block, a critical enhancement not usually focused on in traditional CNN designs. Testing on two extensive public datasets validates the method's strength.

Olsson et al. [22] proposed leveraging deep learning to automatically classify the severity of KOA in adults, using a vast set of unfiltered radiographic knee exams. Utilizing a 35-layer Convolutional Neural Network (CNN) based on the ResNet framework, the team trained the model on 6103 manually labeled images according to the KL scale. Initially trained for 100 epochs without noise interference, followed by 50 epochs with introduced noise factors, the CNN proved highly efficient, achieving an AUC above 0.87 for most KL grades. The study highlights the CNN's robustness in accurately determining knee OA severity even with unclean data, underlining its proficiency and the encouraging potential for medical diagnostic applications.

Chaugule et al. [7] introduced a Deep Convolutional Neural Network (DCNN) model adept at classifying the severity of KOA using digital X-ray images. The model encompasses four stages: autoencoder-driven denoising, segmentation of the image into specific regions, comprehensive feature extraction (comprising region, Zernike, wavelet, and Haralick features), and a feature fusion step to bolster representation. Notably, the DCNN is optimized using the Adam algorithm, registering impressive testing and validation accuracies of 96.31% and 95.70%, respectively. This proposed methodology surpasses other contemporary techniques in discerning KOA severity. An emphasis on the significance of feature extraction, fusion, and pinpointing pivotal features for KOA classification is also presented. Furthermore, the research indicates potential refinements to amplify computational speed and classification efficiency.

Qadir et al. [23] presents an enhanced deep learning algorithm, grounded on a Bidirectional Long Short-Term Memory (BiLSTM) network, tailored for detecting and classifying KOA severity. Employing segmented knee images, features are extracted using ResNet, with the model trained on the Mendeley VI dataset and cross-validated on the Osteoarthritis Initiative (OAI) dataset. The algorithm notably achieves a cross-validation accuracy of 78.57% and a testing accuracy of 84.09%. Assessment utilizes metrics such as recall, accuracy, precision, and F1 score. The system impressively classifies knee images into five categories—Healthy, Phase I to IV—with an overall accuracy of 84.09%, a precision of 92.5%, recall of 99.11%, and an F1 score

TABLE 2. Prior models and their accuracy.

Articles	Dataset	Models	Classification Types	Accuracy
M. Ganesh Kumar et al. [17]	Osteoarthritis Initiative (OAI) (Total image = 4130)	CNN Inception ResNet v2	1. Heathy 2. Unlikely 3. Minimal 4. Moderate 5. Severe	91.03% (CNN Inception ResNet v2)
A. S. Mohammed et al. [18]	1. Osteoarthritis Initiative (OAI) 2. Kaggle (Total image = 9786)	VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNetV2, and DenseNet121	1. Healthy 2. Doubtful 3. Minimal 4. Moderate 5. Severe	Dataset I - 69% (ResNet101), Dataset II - 83% (ResNet101), Dataset III - 89% (ResNet101)
S. S. Abdullah et al. [19]	Radiological center (KGS scancenter, Madurai) (Total image = 3172)	Faster RCNN, ResNet 50, and AlexNet	1. Normal 2. Doubtful 3. Mild 4. Moderate 5. Severe	98.90% (severity classification)
S. A. El-Ghany et al. [20]	Osteoarthritis Initiative (OAI)'s baseline cohort (Total = 8,891 radiographs)	DenseNet169	1. Healthy 2. Moderate 3. Severe	93.78% (Binary classification) 95.93% (Multi-class classification)
T. Tariq et al. [21]	Osteoarthritis Initiative (OAI) (Total image = 9786)	ResNet-34, VGG-19, DenseNet121, and DenseNet161	1. Normal 2. Doubtful 3. Mild 4. Moderate 5. Severe	98% (ENSEMBLE model)
Y. Nasser et al. [13]	1. Osteoarthritis Initiative (OAI) 2. Multicenter Osteoarthritis Study (MOST) (Total image = 7,811 radiographs)	DST-CNN network, DenseNet-121, ResNet-50, Xception, EfficientNet, and MobileNet	1. Normal 2. Doubtful 3. Mild	68.70%. (DST-DNet)
S. Olsson et al. [22]	Osteoarthritis Initiative (OAI) (Total image = 6403)	Convolutional neural network (CNN) of ResNet architecture.	1. Normal 2. Doubtful 3. Severe	87% (CNN)
S. V Chaugule et al. [7]	Mendeley (Total image = 5,778)	Deep CNN	1. Normal 2. Doubtful 3. Mild 4. Moderate 5. Severe	96.31% (DCNN)
A. Qadir [23]	Mendeley VI (Total image = 2,000)	ResNet.	1. Normal 2. Doubtful 3. Mild 4. Moderate 5. Severe	84.09% (Improved ResNet-18)
U. Yunus et al. [24]	Mendeley (Total image = 3,795)	Open exchange neural network (ONNX) and YOLOv2	1. Normal 2. Doubtful 3. Mild 4. Moderate 5. Severe	90.6% (their proposed model)

of 95.69%. Significantly, this approach surpasses current deep learning models in accuracy, robustness, and training and testing durations. Yunus et al. [24] presented a novel method for classifying and pinpointing KOA using radiographic images. This involves transforming two-dimensional radiographs into a three-dimensional format and harnessing LBP features. With the aid of PCA, optimal features are selected, while additional deep features are extracted via AlexNet and Darknet-53 models.

These culminated features, once fused, enable classifiers to reach an accuracy of 90.6% in distinguishing KOA grades. Furthermore, by merging an open exchange neural network (ONNX) with YOLOv2, the system can localize classified images with a mean average precision (mAP) of 0.98. The problem of classifying KOA using deep learning has been highlighted in several studies. This work overcomes these challenges by combining datasets, using sophisticated image pre-processing techniques, and developing a specific model for various knee osteoarthritis classifications. Table 2 presents a comprehensive comparison of the existing related works and Table 3 shows the limitation of those existing related works.

III. METHODOLOGY

Our methodology begins with dataset preparation (Step-1), integrating four distinct datasets to create a comprehensive dataset. Following this, we utilize Generative Adversarial Networks (GANs) for data augmentation (Step-2), enriching our dataset with varied and robust examples to enhance model training. Step-3 involves extensive image pre-processing methods, including a conversion of color channel, applying AHE and FNLM, and resizing function to the augmented images. Then, we proceed to the data splitting method (Step-4) to divide our dataset into training, validation, and testing subsets (for transfer learning models). The subsequent step involves establishing the foundational model (Step-5), which includes adjusting the internal architecture of the CCT model.

Following this, we enact our proposed approach (Step-6), implementing ten distinct modifications to the base models to craft an improved and fine-tuned model designed to meet our specific requirements. Finally, in the results and discussion section (Step-7), we meticulously analyzed the outcomes, evaluating our models using various performance metrics, assessing their robustness with image reduction, and delving into the proposed model's behavior through confusion matrices. This comprehensive and detailed approach ensures a thorough understanding and evaluation of our proposed model compared to existing solutions, highlighting its strengths and areas for potential improvement. Figure 2 depicts the main workflow of our work.

A. DATASET DESCRIPTION

We procured the data from four accessible datasets, namely the KOA Severity Grading and Digital Knee X-ray Images collected from Mendeley: Mendeley I [25] and Mendeley II [26], CGMH KOA Images is collected from Kaggle [27],

TABLE 3. Limitations of earlier studies.

References	Combined Dataset.	Data Augmentation	Apply image Pre-processing Techniques.	Model Fine-tuning	Experimentation with any Transfer Learning Models.
Kumar et al. [17]	N/A	N/A	N/A	N/A	YES
A. S. Mohammed et al. [18]	YES	N/A	N/A	N/A	N/A
S. S. Abdullah et al. [19]	N/A	N/A	N/A	N/A	YES
S. A. El-Ghany et al. [20]	N/A	YES	YES	YES	YES
T. Tariq et al. [21]	YES	N/A	N/A	N/A	N/A
Y. Nasser et al. [13]	N/A	N/A	N/A	N/A	YES
S. Olsson et al. [22]	YES	N/A	YES	N/A	N/A
S. V Chaugule et al. [7]	YES	N/A	YES	N/A	N/A
A. Qadir [23]	N/A	N/A	YES	N/A	N/A
U. Yunus et al. [24]	N/A	N/A	N/A	YES	YES

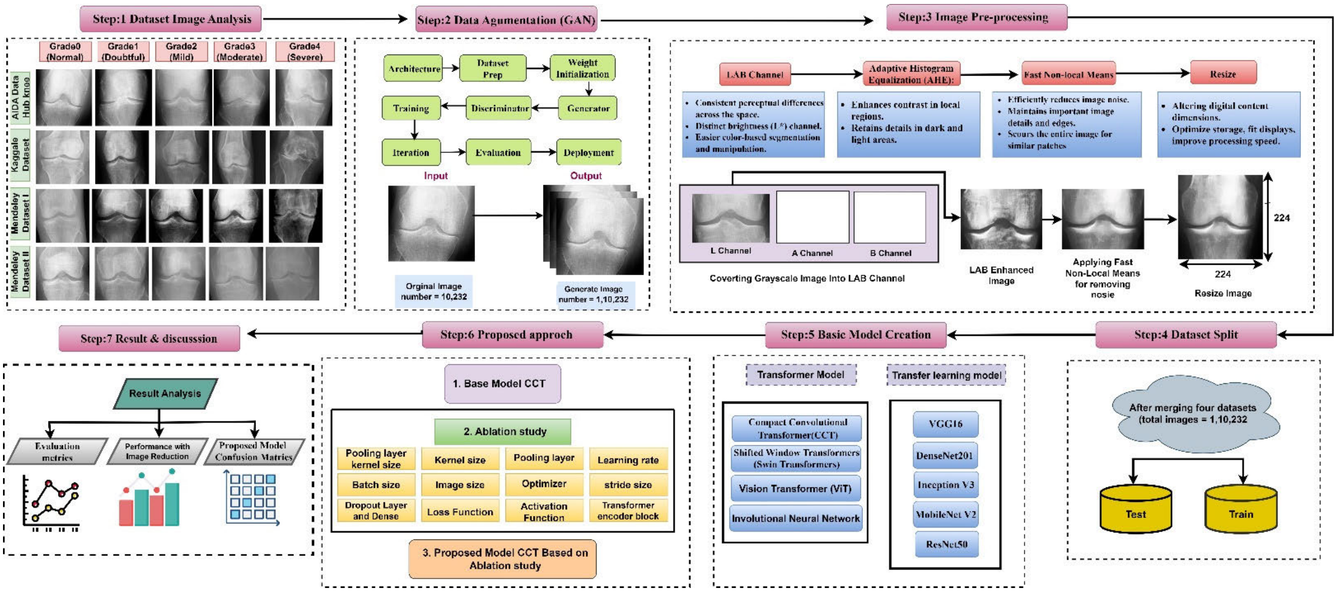

Step:1 Dataset Image Analysis
Grade0 (Normal)
Grade1 (Doubtful)
Grade2 (Mild)
Grade3 (Moderate)
Grade4 (Severe)
AIDA Data Hub knee
Kaggale Dataset
Mendeley Dataset I
Mendeley Dataset II
Step:2 Data Agumentation (GAN)
Architecture
Dataset Prep
Weight Initialization
Training
Discriminator
Generator
Iteration
Evaluation
Deployment
Input
Output
Orginal Image number = 10,232
Generate Image number = 1,10,232
Step:3 Image Pre-processing
LAB Channel
Adaptive Histogram Equalization (AHE):
Fast Non-local Means
Resize
Consistent perceptual differences across the space.
Distinct brightness (L*) channel.
Easier color-based segmentation and manipulation.
Enhances contrast in local regions.
Retains details in dark and light areas.
Efficiently reduces image noise.
Maintains important image details and edges.
Scours the entire image for similar patches
Altering digital content dimensions.
Optimize storage, fit displays, improve processing speed.
L Channel
A Channel
B Channel
Coverting Grayscale Image Into LAB Channel
LAB Enhanced Image
Applying Fast Non-Local Means for removing nosie
224
224
Resize Image
Step:4 Dataset Split
After merging four datasets (total images = 1,10,232
Test
Train
Step:5 Basic Model Creation
Transformer Model
Transfer learning model
Compact Convolutional Transformer(CCT)
Shifted Window Transformers (Swin Transformers)
Vision Transformer (ViT)
Involutional Neural Network
VGG16
DenseNet201
Inception V3
MobileNet V2
ResNet50
Step:6 Proposed approch
1. Base Model CCT
2. Ablation study
Pooling layer kernel size
Kernel size
Pooling layer
Learning rate
Batch size
Image size
Optimizer
stride size
Dropout Layer and Dense
Loss Function
Activation Function
Transformer encoder block
3. Proposed Model CCT Based on Ablation study
Step:7 Result & discusssion
Result Analysis
Evaluation metrics
Performance with Image Reduction
Proposed Model Confusion Matries

FIGURE 2. Workflow diagram.

KOA classification, according to KL is collected from AIDA Data Hub [28]. Each dataset encompasses a range of five distinct grades, which include normal (Grade 0), doubtful (Grade 1), mild (Grade 2), moderate (Grade 3), and severe (Grade 4). The dataset of Digital Knee X-ray images comprises 1650 digital X-ray images of the knee joint, meticulously collected from reputable hospitals and diagnostic centers. Two medical experts have meticulously labeled each radiographic knee X-ray image with KL grades. Furthermore, a pioneering technique has been developed to automatically isolate the cartilage region, which is the region of interest (ROI), based on pixel density [29]. On the other hand, the Mendeley dataset is known as "KOA Severity Grading". This dataset includes knee X-ray data that can be used for both knee joint detection and knee KL grading. After integrating the four datasets, we accumulated 11,431 raw images. Figure 3 shows four distinct types of public datasets and images were captured using an X-ray machine, and the original images are in 8-bit grayscale.

Table 4 represents the distribution of data across various stages of a category, with the data being sourced from four different public datasets: Mendeley Dataset (including a specific subset of 1650 images), Kaggle Dataset, AIDA Data Hub, and an integrated dataset combining four sources. In the 'Normal' stage, there are 3,253 instances from the first Mendeley Dataset, 503 from the second Mendeley Dataset, 80 in the Kaggle Dataset, 248 from AIDA Data Hub, and 4,084 in the integrated dataset. The 'Doubtful' category has 1,495 instances from the first Mendeley Dataset, 488 from the second, 80 from Kaggle, 89 from AIDA Data Hub, and 2,152 in the integrated dataset. For the 'Mild' stage, there are 2,175 instances in the first Mendeley Dataset, 232 in the second, 80 in Kaggle, 89 in AIDA Data Hub, and 2,576 in the integrated dataset. The 'Moderate' stage includes 1,086

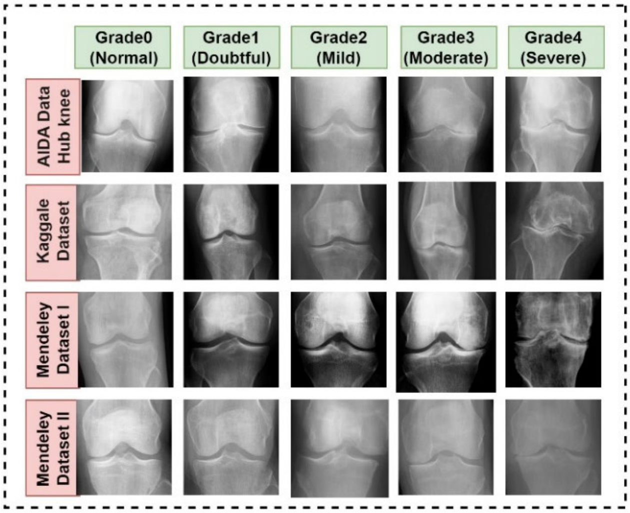


FIGURE 3. Four type of different public dataset.

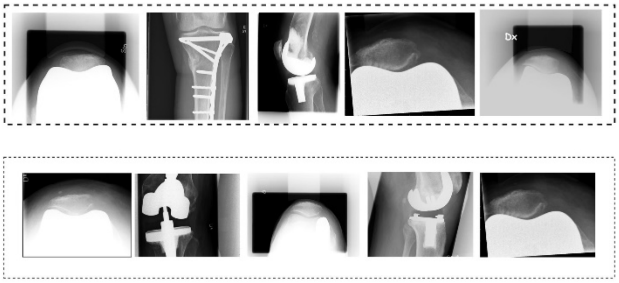


FIGURE 4. Eliminated images from the datahub.

instances from the first Mendeley Dataset, 221 from the second, 80 from Kaggle, 425 from AIDA Data Hub, and 1,812 in the integrated dataset. Lastly, the 'Severe' category has 251 instances from the first Mendeley Dataset, 206 from the second, 80 from Kaggle, 270 from AIDA Data Hub, and 807 in the integrated dataset.

However, we did not include all images in this dataset because some X-ray images were also poor quality and damaged. That is why we removed 1,199 sample images from 11431 images. As a result, our total data set comprises around 10,232 images. Some sample images from the damaged dataset images is shown in Figure 4.

B. EXPERIMENTAL SETUP

After generating a new dataset of 10,232 images, we proceeded with further processing and experimental setup. All experiments were conducted on a system powered by an AMD Ryzen 5 5600X 6-core CPU and 16 GB of RAM. The system was equipped with a ZOTAC GAMING GeForce RTX 3060 Twin Edge OC GDDR6, which claims 12 GB of VRAM. We implemented all models in this experiment using Python 3.9, Keras, TensorFlow v2.15.0, and the PyTorch v1.12.0 framework. We trained a total of 200 epochs in this experiment, and Tables 12 and 13 list all of our experiment's epoch times.

TABLE 4. Knee dataset description.

Stage Name	Mendeley I Dataset	Mendeley II Dataset	Kaggle Data-set	AIDA Data Hub	Integrating four datasets
Normal	3,253	503	80	248	4084
Doubtful	1,495	488	80	89	2152
Mild	2,175	232	80	89	2576
Moderate	1,086	221	80	425	1812
Severe	251	206	80	270	807
Total	8260	1650	400	1121	11,431

C. DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORK

DCGAN can produce stunning, detailed images that resemble actual pictures by combining cutting-edge technology such as deep learning and CNNs. The generator and the discriminator are the two components of DCGAN [47]. While the discriminator's responsibility is to ascertain if a picture is produced or real, the generator's task is to produce new images. These two elements collaborate in a creative dance that continually pushes and tests one another. The discriminator grows better at recognizing the difference between actual and created images as the generator learns and improves at producing images. DCGANs were added to synthetic images to oversample the dataset by enhancing its ability to distinguish between real and imitation samples by maximizing a similar loss function. Due to this, we used DCGAN, which modifies the architecture to ensure stability when combining GAN with deep CNN [30]. Figure 5 depicts the architecture of DCGAN. A discriminative network and a generated network are the two basic types of networks that make up the overall network structure.

They combine their layers into four layers. The generator R that can convert the noise vector z into the sample x is what we are trying to train. A discriminator S. that distinguished between the produced data $pz(z)$ and the real sample data $Pdata(x)$ determined the generator R training objective. The discriminator S will be misled by the generator R into believing that the created data is accurate.

R and S will finally be guided through training to strike a balance in a non-convex game. We employ the gradient descent optimization method without making any prior assumptions or model demands for the data distribution. To train the generator and discriminator networks, use equation (1) [31].

$$V\left(R,S\right) = Ex \sim Pdata\left(x\right)\left[logS\left(x\right)\right] + Ez \sim pz(z)[log(1 - S(R(z)))] \quad (1)$$

The generator uses four convolution2D transpositions and one conv2D layer to sample an image size representation from 14 × 14 × 512 to 224 × 224 × 3. The vector is fed into the dense layer and reshaped to 14 × 14 × 512. Data with a size of 14 × 14 × 512 is transformed into an image with a size

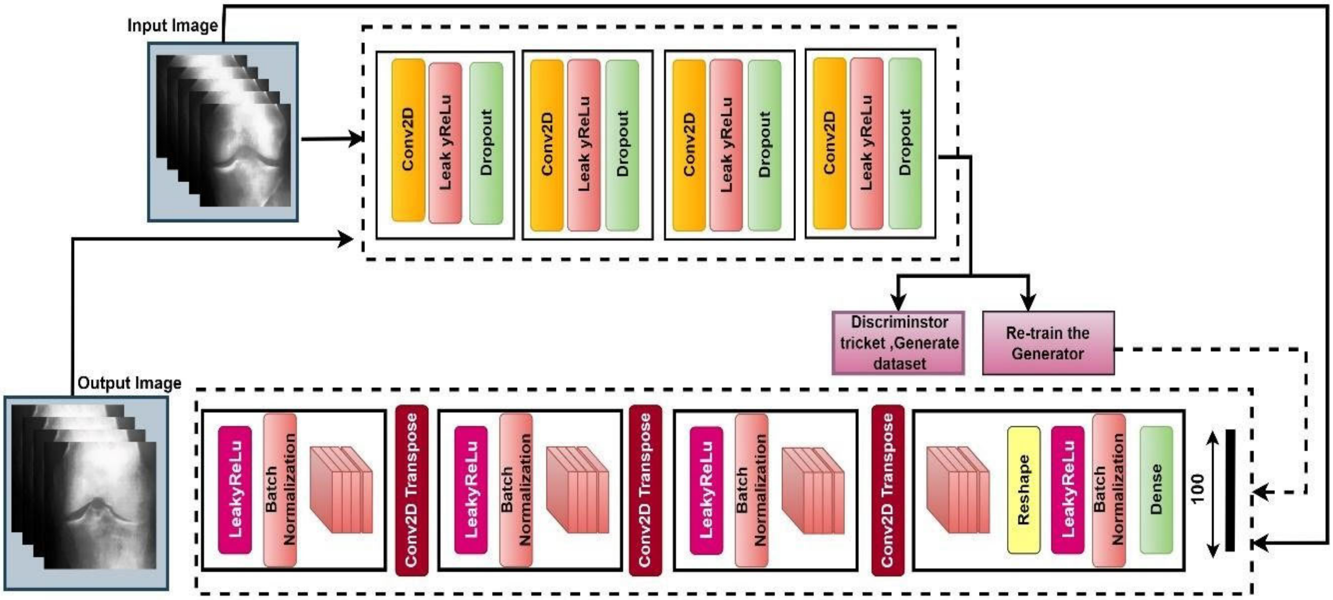


FIGURE 5. Architecture of DCGAN.

of 28 × 28 × 256. The output of the first Conv2D transpose layer is sent via the batch normalization layer, activation function LeakyReLu, and the Conv2D transpose layer before being reshaped to 56 × 56 × 128, 112 × 112 × 64, and 224 × 224 × 32. The Conv2D layer is used in the final layer to provide an output with an image size of 224 ×224 ×3. Figure 6 displays some images generated by DCGAN.

In Table 5, we present a comprehensive dataset combining original and generated images facilitated by DCGAN. Across various grades, ranging from 'Normal' to 'Severe', the total count of images (both original and generated) amounts to 110,232. Specifically, for Grade 0 - Normal, there are 23,782 images; Grade 1 - Doubtful has 22,056; Grade 2 - Mild includes 22,441; Grade 3 - Moderate consists of 21,408; and Grade 4 - Severe contains 20,545 images. When we delve into the distribution of original images within these grades, Grade 0 - Normal comprises 3,782 images, Grade 1 - Doubtful has 2,056, Grade 2 - Mild includes 2,441, Grade 3 - Moderate has 1,408, and Grade 4 - Severe contains a significant 545 images, with a total of c original images across all grades. This meticulous compilation and augmentation of the dataset using DCGAN ensure a robust and diverse set of images, facilitating enhanced performance and reliability in subsequent analyses or model training processes.

The figure 7 compares a DCGAN-generated image with an original medical image. Anatomically, both images show a comparable structure, most likely a knee joint. The bottom left displays a histogram that shows the image's intensity distribution, while the top left displays the original image. A few of the original image's most important statistical features are as follows: The mode is 173 with a count of 3403, the total count is 1576, the maximum intensity is 255, the minimum intensity is 1, the standard deviation is 49.559, and the value is 196. The top right corner displays the DCGAN-generated picture, while the bottom right displays the matching histogram. The resulting image has a value of 194, a count of 3804, a total count of 1832, a minimum intensity of 0, a mean intensity of 143.546, a maximum intensity of 253, and a standard deviation of 52.995. The histograms illustrate variations in mean intensity, standard deviation, and mode by displaying the distribution of pixel intensity values in the two images. The original images mean intensity is marginally greater than the produced images. In addition, a greater standard deviation in the resulting image indicates a broader distribution of intensity values. The mode values indicate a difference in the most common intensity values between the two images, and both have comparable maximum intensity values that are near the upper limit of 255. This comparison shows how well DCGAN performs in producing images that are almost identical to the original, despite minor differences in statistical characteristics and intensity distribution.

D. IMAGE PRE-PROCESSING

Image pre-processing is crucial for ensuring optimal computation time and enhanced model performance, and it must be completed before inputting the images into a neural network. In this phase, we started by cropping the selected portion of the image. Following that, we performed a Lightness, channel a and channel b (Lab) color space transformation and applied the AHE is utilized to enhance the contrast of the images, which is particularly beneficial for images with low contrast or uneven illumination, as it redistributes the pixel values to make the image details more discernible. To further refine the image quality, we removed noise using FNLM. filter by applying to the L channel of the LAB color space [32]. This process involves comparing each pixel to its nearby pixels and averaging their values based on their similarity, resulting in a smoother image. Lastly, we resized to a consistent resolution

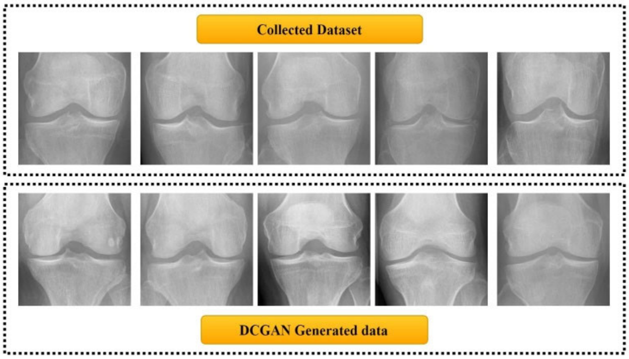


FIGURE 6. Original data and DCGAN Generated Data.

TABLE 5. Number of original and generated data using DCGAN.

Stage Name	Total image
Grade 0- Normal	23782
Grade 1-Doubtful	22056
Grade 2-Mild	22441
Grade 3-Moderate	21408
Grade 4-Severe	20545
Total	**1,10,232**

of 224 × 224 pixels with FNLM for subsequent use. However, Figure 8 shows the total image processing workflow and all the methods are demonstrated below.

1) ADAPTIVE HISTOGRAM EQUALIZATION (AHE)

The AHE in medical imaging primarily aims to enhance image contrast, particularly in regions crucial for diagnosis. Medical images often present interpretation challenges due to their inherently poor contrast, complicating the differentiation between various tissue types and identifying potential abnormalities. AHE tackles this issue by adaptively redistributing intensity values across the image, focusing on local regions instead of the entire image [33]. This localized approach to contrast enhancement ensures that subtle details and vital diagnostic features are more visible and pronounced, enabling medical professionals to interpret the images more accurately and reliably. Ultimately, AHE contributes to precise diagnosis and treatment, improving patient outcomes. Although AHE can be complex to implement, it has proven particularly effective for medical images, as evidenced by several studies [34]. Some methods have simplified AHE's application, such as using fewer pixels or adjusting the surrounding area of each pixel [35]. In the context of X-ray imaging, AHE enhances contrast by dividing the image into small sections called tiles, each of which is adjusted individually to highlight its details. This adaptive approach ensures bright and dark areas are distinctly visible, making identifying crucial bodily features and any existing health issues easier. The output after employing AHE method is shown in Figure 8.

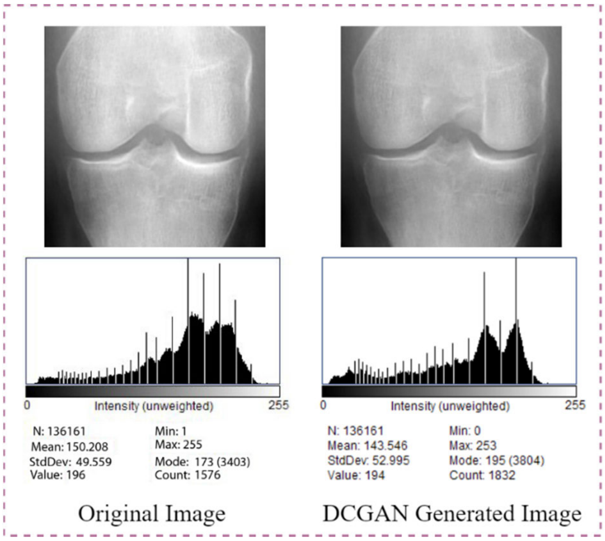


FIGURE 7. Original image and DCGAN generated images histogram visualization.

2) FAST NON-LOCAL MEANS

Utilizing the FNLM for X-ray images significantly enhances image quality by effectively reducing noise while preserving crucial structural details. This advanced denoising technique outperforms many traditional methods, particularly in keeping edges and fine details, which are vital for accurate diagnosis and analysis in medical imaging. The algorithm achieves this by considering a wider range of pixels for denoising, extending beyond local neighborhoods to capture the inherent patterns and structures in the data. As a result, medical professionals benefit from more precise, more reliable images, leading to improved detection and characterization of abnormalities [36]. Additionally, the 'fast' variant of Non-Local Means (NLM) ensures quicker processing times, making it more practical for clinical settings where time is of the essence. Ultimately, FNLM contributes to more accurate diagnostic decisions and potentially better patient outcomes. The FNLM is a quick and efficient way to clean up noisy images. It is based on the NLM method, which studies the entire image and finds patterns to reduce noise. However, NLM can be a bit slow, especially with large images. So, FNLM was created to speed things up [37]. It uses some clever shortcuts and tools to find and fix noise faster. While it might not be as thorough as the original NLM in some cases, FNLM strikes a good balance between speed and quality, making it handy for on-the-fly edits or handling big batches of images. The noisy image is represented by $Y = X + N$. and denoised pixel value $X(i)$ is given by

$$\boldsymbol{X}(i) = \frac{1}{Z}\sum\nolimits_{j\epsilon I}^{n} \omega_{ij} Y(i) \tag{2}$$

where, w_{ij} is the weight denoting the contribution from pixel $Y\ (i)$ to the denoised pixel $X(i)$.

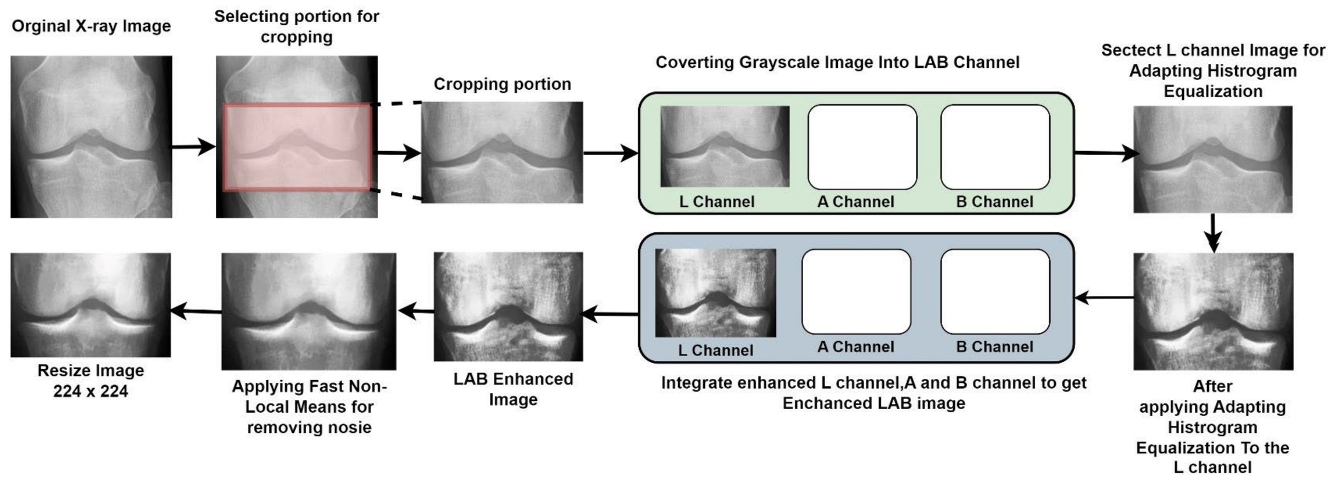

Orginal X-ray Image
Selecting portion for cropping
Cropping portion
Coverting Grayscale Image Into LAB Channel
L Channel
A Channel
B Channel
Sectect L channel Image for Adapting Histrogram Equalization
After applying Adapting Histrogram Equalization To the L channel
L Channel
A Channel
B Channel
Integrate enhanced L channel,A and B channel to get Enchanced LAB image
LAB Enhanced Image
Applying Fast Non-Local Means for removing nosie
Resize Image 224 x 224

FIGURE 8. Image pre-processing workflow.

3) VERIFICATION

A statistical analysis was conducted to show that the algorithms do not negatively affect image quality. The equations for various verification techniques are listed below.

MSE is perhaps the most often used and simple loss function. MSE is computed by taking the square of the difference between the actual data and the model's predictions and then averaging the result for the whole dataset. The following equation provides the mathematical definition of MSE:

$$MSE = \frac{1}{pq}\sum\nolimits_{i=0}^{x-1}\sum\nolimits_{i=0}^{y-1}(O(x,y),-P(x,y))^2 \quad (3)$$

Here p and q denote the pixels of O and P, x and y denote the rows of the pixels p and q, where O is the original picture and P is the processed image. A number around 0 denotes high picture quality. The MSE value goes from 0 to 1.

In this study, we used PSNR to calculate the signal-to-noise ratio and to compare the quality of a picture between its original and compressed versions. With increasing PSNR, the image quality improves. The following equation provides the mathematical definition of PSNR:

$$PSNR = 20log_{10}(\frac{(MAX)}{\sqrt{MSE}}) \quad (4)$$

The maximum pixel value of the picture is indicated here by MAX. An 8-bit picture should typically have a PSNR of between 30 and 50 dB.

A statistic called SSIM quantifies the loss of picture quality brought on by image processing. A reference picture and a processed image with the same image origin are required. The SSIM equation is given by

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (5)$$

Model predictions or estimations are routinely compared with actual observed data using the root mean square error (RMSE). Table 6 shows that MSE values greater than 0.33, PSNR values greater than 31, SSIM values greater than 0.99, and RMSE values greater than 0.54 ensure that our preprocessed image has good quality.

E. TRANSFER LEARNING MODELS

After applying our image pre-processing techniques, we conducted tests using five pre-trained models: VGG16, DenseNet201, InceptionV3, MobileNetV2, and ResNet50. The following sections briefly describe each pre-trained model.

1) VGG16

VGG16 model consists of 16 weighted layers, mainly using 3 × 3 convolutional layers, making it both deep and relatively straightforward [38]. VGG16 excelled in the ImageNet challenge, a key benchmark in image classification. While its depth gives it strength, its high computational cost has paved the way for more efficient models like ResNet. Nevertheless, VGG16 remains a foundational model in deep learning, valuable for various applications beyond more image classification.

2) DenseNet201

Its unusual densely linked topology distinguishes a deep convolutional neural network with 201 layers, DenseNet201. DenseNet201 boasts a novel architecture where each layer receives inputs from all of its preceding levels, in contrast to typical networks where layers are connected sequentially [39]. Due to the effective transmission of features and gradients across the network made possible by this dense connection design, learning and feature reuse are increased. Despite its depth, DenseNet201 achieves outstanding parameter efficiency, albeit at the expense of higher memory usage [40]. DenseNet201, which is at the cutting edge of deep learning architecture developments, emphasizes the possibility of enhancing intra-network connection to improve performance in challenging image classification tasks.

TABLE 6. MSE, PSNR, SSIM, RMSE value for seven images.

Image	MSE	PSNR	SSIM	RMSE
Image 1	0.43	31.55	0.9923	0.68
Image 2	0.40	32.20	0.9955	0.63
Image 3	0.41	32.08	0.9946	0.61
Image 4	0.42	31.96	0.9935	0.64
Image 5	0.31	33.23	0.9951	0.55
Image 6	0.34	32.85	0.9917	0.58
Image_7	0.33	32.74	0.9945	0.55

3) InceptionV3

This model has higher performance in object detection and has three distinct parts: the initial convolutional block, the classifier, and the improved inception module. In this model, for accelerating the training speed and reducing the number of feature channels, a 1 × 1 [41] convolutional kernel is highly used. This model is built with different layers, such as max-pooling layers, average pooling layers, conventional layers, dropout layers, and fully connected layers. And then, to show the result, the SoftMax activation function is associated with a fully connected (FC) layer.

4) MobileNetV2

MobileNetV2 stands out in the deep learning as a compact and efficient model designed with a mobile-first approach to balance performance and efficiency [42]. It features an inverted residual structure and linear bottleneck, significantly reducing size and complexity while maintaining essential information. This makes it suitable for mobile and embedded vision applications, offering strong performance in image classification and recognition tasks, even in scenarios with limited computational resources or power [43].

5) ResNet50

Another CNN architecture is ResNet50. In this architecture, the model contains 50 layers. In this model, there is a shortcut route for reaching the final state, and during the training period, this shortcut route helps avoid the unusual layers [41], making the entire process faster. This model has over 23 million parameters.

F. TRANSFORMER MODELS

In this study, we also trained several transformer models, including the compact convolutional transformer (CCT), vision transformer (ViT), shifted window transformer (Swin), and the Involutional neural network. We applied the enhanced dataset to each model to determine our research's most suitable base model.

1) SHIFTED WINDOW TRANSFORMERS (SWIN TRANSFORMERS)

The Swin Transformer, commonly termed as Swin is an innovative structure in the computer vision domain. Essentially, it breaks down an image into smaller segments or patches. Its method of handling these patches is unique to the Swin: as they delve deeper into the transformer layers and are merged to generate a layered hierarchical understanding of the image [44]. Swin starts by partitioning an image into non-overlapping patches. These patches are then linearly embedded. While typical transformers would look at all patches, Swin smartly chooses a subset of adjacent patches and applies self-attention locally within this window. To prevent edge patches from only attending to a limited set of neighbors, swin rotates or ''shifts'' these windows in subsequent layers. The transformer layers are organized hierarchically. As progress progresses deeper, adjacent patches are merged, effectively increasing the receptive field without significantly adding to the computational cost [45]. Swin can be paired with architectures like Feature Pyramid Networks (FPN) or U-Net to refine its dense predictions further. One of the standout features of Swin is that its computational complexity grows linearly with the image size. This starkly contrasts traditional transformers, where the complexity can increase quadratically. Using these architectural strategies, the Swin Transformer balances local and global self-attention, making it a versatile and efficient backbone for many visual tasks.

$$Attention(Q, K, V) = Softtmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

where:

- *Q, K, and V are the query, key, value matrices.*
- *dk is the dimension of the keys.*

2) COMPACT CONVOLUTIONAL TRANSFORMER (CCT)

Compact Convolutional Transformer (CCT) model is an innovative blend of convolutional neural networks (CNNs) and transformer architecture [46]. In simple terms, CCT uses filters like CNNs to break down images and then uses transformers to understand them better. This combined structure makes it efficient and effective, especially for diverse image data. The CCT uses convolutional layers to process input images at the onset [47]. These layers are adept at extracting intricate spatial features from the images. They can recognize patterns, edges, and other essential elements, making them the foundation of many image-related tasks. After the convolutional processing, the extracted features undergo tokenization. This step transforms the spatial features into a sequence of tokens. These tokens represent parts of the image in a format that transformers can understand and manipulate. Post-tokenization, these sequences are fed into the transformer architecture. Unlike traditional CNNs, transformers are particularly skilled at capturing long-range dependencies and relationships between different parts of an image [48]. The self-attention mechanism in transformers allows each token to weigh its relationship with every other token, leading to a more comprehensive understanding of the image's context. After processing through the transformer layers, the model generates an output that can be used for various tasks, such as classification, segmentation, or any other relevant objective. In this method, we use convolutional blocks to extract portions of the image, known as patches.

$$Image(Y) \in RH\ x\ W\ x^{C} \tag{7}$$

$$\mathrm{Yo} = \mathrm{MaxPool}(ReLU(Conv2D(Y))) \tag{8}$$

For any image Y with height (H), width (W), and channels (C), these patches are then turned into a sequence of a certain length (l). In Figure 9, shows the Base CCT model architecture.

3) VISION TRANSFORMER (ViT)

Convolutional Neural Networks (CNNs) have dominated the landscape of image-processing tasks. Inspired by the human visual system, these networks utilize convolutional layers to process spatial hierarchies in images, capturing local features and patterns. However, recent advances in deep learning have introduced a paradigm shift in how we perceive and process visual data. Enter the ViT [49]. Borrowing from the success of transformers in natural language processing tasks, the ViT seeks to apply similar principles to the domain of computer vision. Instead of exploiting spatial hierarchies, the Vision Transformer focuses on a sequence of non-overlapping image patches, representing each as a linear embedding. These patches are then processed in parallel through self-attention mechanisms, enabling the model to capture long-range dependencies and intricate patterns within the image [50]. The promise of ViT lies not just in its novel approach but also in its scalability. Similar large transformers have benefitted natural language tasks, and the Vision Transformer's performance improves with increased data and computational resources. In many benchmark tasks, ViTs, especially when pre-trained on vast datasets, have either matched or surpassed traditional CNN-based approaches.

4) INVOLUTIONAL NEURAL NETWORK

The Involutional Neural Network is a unique twist on traditional neural structures, focusing on capturing local and global contextual information in a novel manner. Unlike the consistent kernel application in typical convolutional networks, involution generates dynamic kernels for each pixel based on spatial location [51]. This involves a kernel generation layer that uses operations like point-wise convolution. To optimize computational efficiency, the architecture incorporates channel reduction and splitting techniques [52]. This ensures the network remains adaptive to spatial variations while managing computational demands, creating a blend of precision and efficiency.

G. PERFORMANCE OF OUR PROPOSED MODEL

Our research primarily aimed to refine the KOA classification technique through X-ray imagery. We performed nine ablation studies to determine the best-performing setup for our model.

1) BASE COMPACT CONVOLUTIONAL TRANSFORMER (CCT) MODEL

In a recent analysis, we compared the performance of four transformer models and found that the CCT model was highly effective. By performing ablation investigations on a base CCT model, a modified version of the model is proposed in this study. Figure 9 presents the architecture of our base CCT model. We saw the potential to improve its performance further and modified its fundamental framework to create a more advanced version. Initially, it takes 32 × 32 ×3 dimension images and enhances them using various geometric augmentations. These improved images are then resized to 36 × 128 dimensions through a process involving the CCT tokenizer and other elements such as a convolution layer and pooling layer with specific technical parameters (like a stride size of 2, a kernel size of 5 for the convolution layer, and a kernel size of 4 for the pooling layer). Following reshaping, the image goes through a series of intricate modifications in two major encoder blocks, each with several layers for functions, including normalization, regulation, and multi-head attention. These blocks clean up the image data, keeping the output in the 36 × 128 dimension during this stage [53]. Following this, the output moves through another normalization layer, transitioning into a dense layer paired with a Softmax layer, further producing the dimensions to 64 × 1. This output is then streamlined to a 1 × 128 dimension through a sequence pooling layer. The X-ray image is then divided into five classes using a linear classification layer that processes the improved data. Table 7 outlines our proposed base CCT model. Where we have a data augmentation layer, which input images of 32 × 32 pixels to improve the model's ability. Then we have CCT Tokenizer transforms the data into a format the model can understand with a size of 36,128. Central to the model are two transformer encoder blocks crucial for interpreting data as shown in figure 9. Layer normalization with a size of 36,128 also helps in achieving more uniform training results. The model includes dense layers, which are types of fully connected neural network layers. For turning raw scores into a distribution of probabilities, the Softmax function is utilized. Additionally, sequence pooling, which has a size of 1,128, allows the model to manage inputs of different sizes.

Each part of this model serves as a fundamental component, with most being used only once, but the encoder blocks are utilized twice. Additionally, Categorical Cross entropy is chosen as the loss function, and the Adam optimizer is applied with a learning rate of 0.001. The model is executed with a batch size of 128 over 200 epochs.

Table 7 presents the details of the transformer block. It has two-layer normalization, two dropout, two dense layer, three regularization layer and a multi head attention layer.

a: RESULTS OF THE MODEL ABLATION

As noted before, we carried out ablation research on the fundamental CCT model to improve performance through

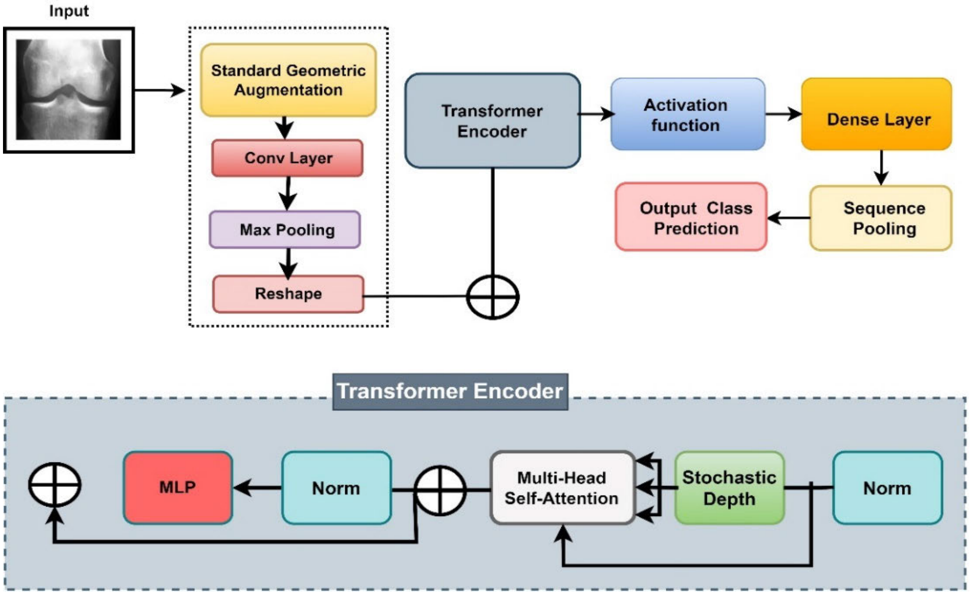


FIGURE 9. CCT model architecture.

TABLE 7. Architecture details of the base CCT model.

Layer	Size	Number of Block
Data Augmentation	32,32,3	1
CCT Tokenizer	36,128	1
Transformer Encoder Block	-	2
Layer Normalization	36,128	1
Dense	-	1
Softmax	-	1
Sequence Pooling	1,128	1
Dense	-	1

TABLE 8. Details of transformer encoder block.

Layer	Size	Number of Block
Layer Normalization	36,128	2
Dropout	-	2
Dense	-	2
Regularization (stochastic depth)	36,128	3
Multi Head attention	36,128	1

layer layout modifications. We conducted nine studies modifying kernel size, stride size, batch size, batch learning rates, various optimizers, and pooling layer kernel size. In our study, we employed an ablation study approach to optimize the Base CCT model. This process involves selecting a standard value for each hyperparameter and conducting experiments with varied parameter values.

We iteratively refine the parameters, using the best-found values as the starting point for subsequent experiments. With this methodical process, we are able to determine the ideal configuration for our suggested CCT model. These ablation studies on the initial CCT model led to a more stable architecture with faster processing speeds and better classification accuracy. The results of these experiments are comprehensively detailed in Table 9 and 10.

Study 1 (Modifying the Transformer Layers): We added or removed several encoded blocks from the base model's transformer layers in order to get the best results. The model worked best with the first configuration, training in 170 seconds for each cycle and getting the highest accuracy. However, the second and third configurations also got good results, with 89.56% and 90.50% accuracy. Since the third configuration took the shortest time to train, we decided to use it for our next set of experiments.

Study 2 (Modifying Dense and Dropout Layers): The quantity of dense and dropout layers may influence a classifier's performance. In this study, we used different combinations of dropout and dense layers. Our basic model had the best accuracy in configuration 1 (90.68%), but it also required the most training time per epoch (122 s). Our base model (configuration 3) follows the addition of dropout and dense layers and achieved the second-highest accuracy with 120 s per epoch. Our model's test accuracy for configuration 2 was 90.30%, and its training time for each epoch was 121s. Although the third configuration achieved accuracy very close to the second highest, training took less time than the other configurations. As a result, configuration 3 was picked for additional testing.

Study 3 (Altering the Activation Function): The efficacy of a classification model is influenced by the activation functions used. Finding the best activation function can improve a model's performance. We also used several additional activation functions, such as the soft sign, soft plus, rectified

TABLE 9. Ablation study on changing transformer encoder block dense layer, dropout layer, activation function, and pooling layer.

Case Study 1: Changing the Transformer Encoder Block				
Configuration No.	No. of transformer encoder blocks	Epoch x time	Test Accuracy (%)	Findings
1	3	200 x 170s	89.61%	Accuracy improved
2	2	200 x 138s	89.56%	Base Model accuracy
3	1	200 x 122s	90.50%	Highest accuracy
Case Study 2: Changing the Dropout Layer and Dense Layer				
Configuration No.	No of dropout layer	No of dense layer	Epoch x Time	Test Accuracy (%)
1	3	3	200 x 122s	90.68%
2	2	2	200 x 121s	90.30%
3	1	1	200 x 120s	90.46%
Case Study 3: Changing the activation function				
Configuration No.	Activation function	Epoch x Time	Test Accuracy (%)	Findings
1	relu	200 x 120s	90.84%	Highest accuracy
2	elu	200 x 120s	87.23%	Accuracy dropped
3	softsign	200 x 120s	85.32%	Accuracy dropped
4	softplus	200 x 120s	84.14%	Accuracy dropped
Case Study 4: Changing the Pooling Layer				
Configuration No.	Image Size	Epoch x time	Test Accuracy (%)	Findings
1	Max	200 x 120s	91.51%	Highest accuracy
2	Average	200 x 120s	90.84%	Previous accuracy

linear unit (ReLU), and exponential linear unit (ELU). Table 9 displays the ReLU activation function's performance, revealing a best test accuracy of 90.84%. Every epoch (120 s) has the same training time for every activation function. In this regard, we chose a future investigation on ReLU activation.

Study 4 (Changing the Pooling Layer): We tested two different pooling layers: maxpooling and average pooling. Both took 120 seconds to train in each epoch. The maxpooling layer came out on top with the highest accuracy, reaching 91.51%. So, we chose to use maxpooling for our next steps.

Study 5 (Altering the Stride Size): In this research, we explored various stride sizes in the transformer layers of the model. We tested sizes 1, 2, 3, and 4, each taking 120 seconds for every training cycle. The first configuration, using a stride size of 1and achieved the highest accuracy (91.51%). This led us to choose stride size 1 for the next steps in our research.

Study 6 (Altering the Kernel Size): We conducted tests using various kernel sizes (4, 3, 2, and 1) in the transformer layers. The tests showed that the 4-sized kernel was the most efficient, yielding a top accuracy of 91.91% and requiring 122 seconds for each training session. While the 3-sized kernel also performed reasonably well with an 85.77% accuracy, it did not match the performance of the 4-sized kernel. Hence, we have selected the 4-sized kernel for in-depth future research.

Study 7 (Altering the Batch Size): Changing the batch size can affect how well the classification works. So, we have used different batch sizes: 256, 128, 64, and 32. The 32-batch size gave us the best accuracy at 91.96%, but it also took a long

TABLE 10. Ablation study on changing stride size, kernel size, batch size, optimizer and learning rate.

Case Study 5: Altering the stride size				
Configuration No.	No. of strides	Epoch x Time	Test Accuracy (%)	Findings
1	1	200 x 120s	91.51%	Previous accuracy
2	2	200 x 120s	89.14%	Accuracy dropped
3	3	200 x 120s	88.22%	Accuracy dropped
4	4	200 x 120s	86.34%	Accuracy dropped
Case Study 6: Altering the Kernel Size				
Configuration No.	No. of Kernel Size	Epoch x Time	Test Accuracy (%)	Findings
1	4	200 x 122s	91.91%	Highest accuracy
2	3	200 x 120s	91.51%	Previous accuracy
3	2	200 x 117s	85.77%	Accuracy dropped
4	1	200 x 115s	83.95%	Accuracy dropped
Case Study 7: Altering the batch size				
Configuration No.	Batch size	Epoch x Time	Test Accuracy (%)	Findings
1	256	200 x 120s	90.25%	Accuracy dropped
2	128	200 x 119s	91.91%	Previous accuracy
3	64	200 x 133s	91.94%	Accuracy improved
4	32	200 x 145s	91.96%	Accuracy improved
Case Study 8: Altering the Optimizer				
Configuration No.	Optimizer	Epoch x Time	Test Accuracy (%)	Findings
1	Adam	200 x 120s	92.20%	Highest accuracy
2	Nadam	200 x 120s	90.18%	Accuracy dropped
3	SGD	200 x 120s	91.91%	Previous accuracy
4	Adamax	200 x 120s	85.44%	Accuracy dropped
5	RMSprop	200 x 120s	87.35%	Accuracy dropped
Case Study 9: Altering the learning rate				
Configuration No.	Learning rate	Epoch x Time	Test Accuracy (%)	Findings
1	0.01	200 x 120s	87.12%	Accuracy dropped
2	0.006	200 x 120s	92.20%	Previous accuracy
3	0.001	200 x 120s	94.38%	Highest accuracy
4	0.0008	200 x 120s	93.11%	Accuracy improved

time 145 seconds for each training session. The 128-batch size took 119 seconds and was almost accurate. That's why we decided to use the 128-batch size for our next studies.

Study 8 (Altering the Optimizer): We have used five optimizers to find the best one: Adam, Nadam, SGD, Adamax, and RMSprop. The Adam optimizer came out on top with the highest accuracy of 92.20%. The other optimizers performed well, as well, but they couldn't beat Adam's score. So, we've chosen to stick with the Adam optimizer for our future tests.

Study 9 (Altering the Learning Rate): We tested various learning rates: 0.01, 0.006, 0.001, and 0.0008 to see which worked best. Everything else stayed the same during these tests. The 0.001 rate gave us the best accuracy, so we used it for our planned model. Table 11 and 12 represent the optimal configurations of the proposed model and model design and tuning, respectively.

2) PROPOSED MODEL

In order to optimize performance while minimizing time complexity and training durations, we strengthen and shorten the proposed CCT architecture. Our based CCT architecture has two transformer encoder blocks, whereas after ablation study, the proposed KOA-CCT model has just one transformer encoder blocks. The resultant CCT architecture is robust and allows for shorter training durations than the original CCT architecture. This version maintains the strengths of the previous one while making it more streamlined and efficient. The configuration of our proposed KOA-CCT model shown in Figure 10 for a clear understanding on a visual level. The KOA-CCTNet model differs from the traditional CCT model principally by its simplified structure. To maintain model efficacy, further adjustments are made, such as reducing the stride and kernel sizes to 1 and 4, respectively. To fit this new framework, we also change the CCT tokenizer's output's size to 64 × 128. This novel model does not require positional encoding, a technique frequently applied in previous transfer-based models. This absence significantly reduces the computational demands, characterized mathematically as reducing the complexity from,

$$O(L^{\wedge}2D + LD^{\wedge}2) \tag{9}$$

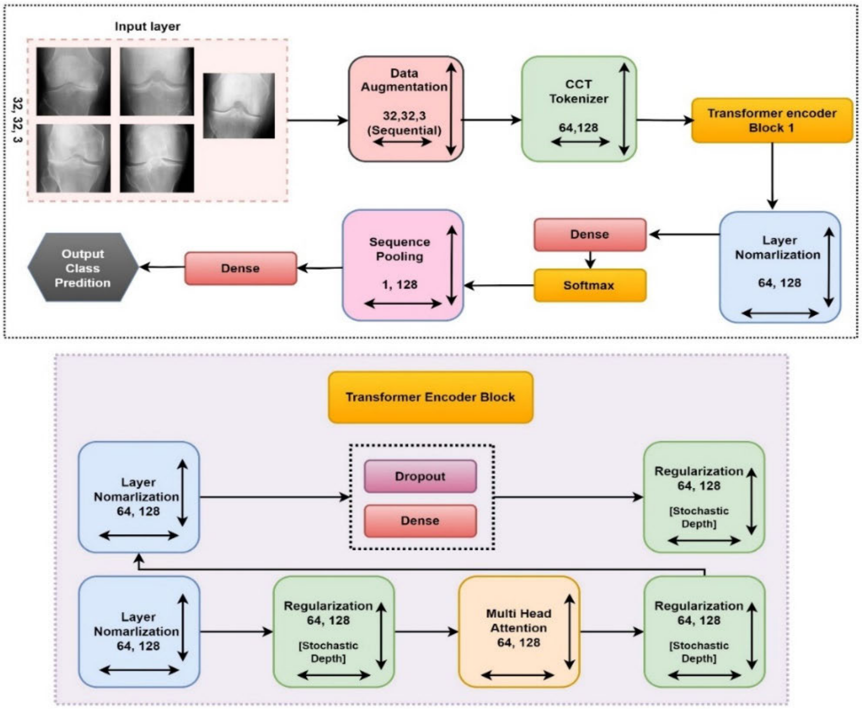


FIGURE 10. Architecture of the proposed KOA-CCTNet model.

TABLE 11. Configuration of proposed model after the ablation study.

Configuration	Value
Image size	32 x 32
Epochs	200
Optimization function	Adam
Learning rate	0.001
Batch size	128
Kernel size	4
Activation function	ReLU
Loss Function	Categorical Cross-Entropy
Kernel size of the pooling layer	4
Stride size	1
Pooling layer	Max pooling
Projection_dim	128
Stochastic_depth_rate	0.1
Weight_decay	0.0001

TABLE 12. Model design and tuning.

Layer (type)	Output Shape	Param #
input_1	[(None, 32, 32, 3)]	0
data augmentation (Sequential)	[(None, 32, 32, 3)]	0
cct_tokenizer (CCTTokenizer)	(None, 64, 128)	134144
Transfer Encoder Block	(None, 64, 128)	0
layer_normalization_1(LayerNormalization)	(None, 64, 128)	256
dense (Dense)	(None, 64, 128)	16512
tf.nn.softmax (TFOpLambda)	(None, 64, 1)	0
dense_1 (Dense)	(None, 5)	645

where D is the dimensionality of the vector representation and L is the length of the input sequence [54]. This trimmed-down approach means that the KOA-CCTNet needs fewer resources to function, speeding up the training and testing phases without compromising performance and enhancing the model's overall efficiency.

IV. RESULTS AND ANALYSIS

A. PERFORMANCE METRICS

We used several metrics such as accuracy, precision, recall, F1-score, specificity, False Positive Rate (FPR), False Negative Rate (FNR), False Discovery Rate (FDR), Negative Predictive Value (NPV), Matthews Correlation Coefficient (MCC) to assess the proposed classification model in this work. The confusion matrix, accuracy and loss curve for the proposed model is also shown in Figure 11C. When the model correctly identifies a positive class, the result is true positive (TP). A true Negative (TN) is an outcome where the model properly categorizes a negative class. When the model incorrectly predicts the positive class, it is called false positive (FP), while false negative is an output in which the negative class is mispredicted. The value or performance metrics were calculated using equations (10)-(19), [52], [53], [54], [55], [56].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F_1 - score = 2\frac{precision \times recall}{precision + recall} \tag{13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

$$FPR = \frac{FP}{FP + TN} \tag{15}$$

$$FNR = \frac{FN}{FN + TP} \tag{16}$$

$$FDR = \frac{FP}{FP + TP} \tag{17}$$

$$NPV = \frac{TN}{TN + FN} \tag{18}$$

$$MCC = \frac{TNXTP - FPXFN}{\sqrt{(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}} \tag{19}$$

B. RESULTS OF THE TRANSFER LEARNING MODELS

In our research, we first experimented with five transfer learning techniques. Table 13 shows that all yielded lower accuracy levels, demanded extended training times, sensitivity, specificity, recall, f1 score and AUC. We tested these techniques using our improved merged dataset. Each model was trained for 200 epochs. While MobileNetV2 achieved the highest accuracy, it was time-consuming. The other models also exhibited significant training times and comparatively lower performance. Nevertheless, there is a chance to improve the performance of the model by reducing the time during classification.

C. RESULTS OF THE TRANSFORMER MODELS

This study examined four transformer models using images from our dataset. Every model ran for 200 epochs. We aimed to find a model that was both accurate and efficient. As per Table 14, the CCT model got 86.54% accuracy in 225 seconds for each epoch, while the ViT model was the second best accurate at 84.81% but took more time. With an accuracy of 83.58%, the involutional neural network had the lowest accuracy, while the Swin transformers was comparable to the ViT's. Given that the CCT model was the fastest, we chose it as the foundation for our ongoing research.

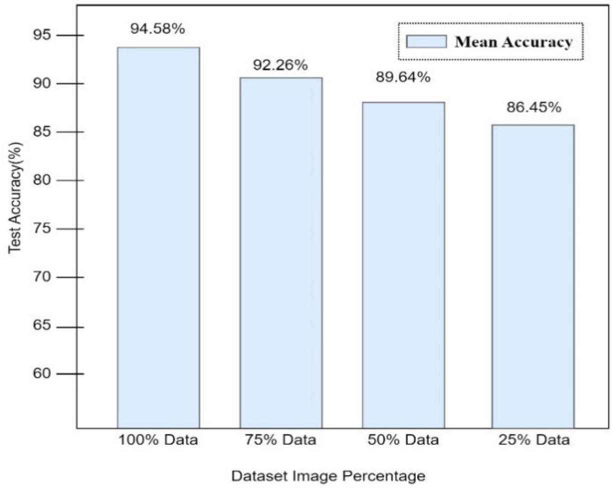


FIGURE 11. Results of image reduction.

TABLE 13. Comparison of performance in transfer learning models considering accuracy and epoch duration.

Transfer Learning Model	Accuracy	sensitivity	specificity	Recall	F1 score	AUC	Epoch × Time
Resnet 50	76.89 %	73.21	83.21	73.78	73.56	73.88	200 x 15375s
VGG16	78.67 %	75.53	83.95	75.88	75.66%	76.33	200 x 12375s
Densnet201	79.98 %	77.83	78.95	77.94	78.33	78.44	200 x 10205s
InceptionV3	80.23 %	78.78	78.89	78.29	78.52	79.21	200 x 1956s
MobileNetv2	80.77 %	78.86	78.92	78.55	79.21	79.56	200 x 1875s

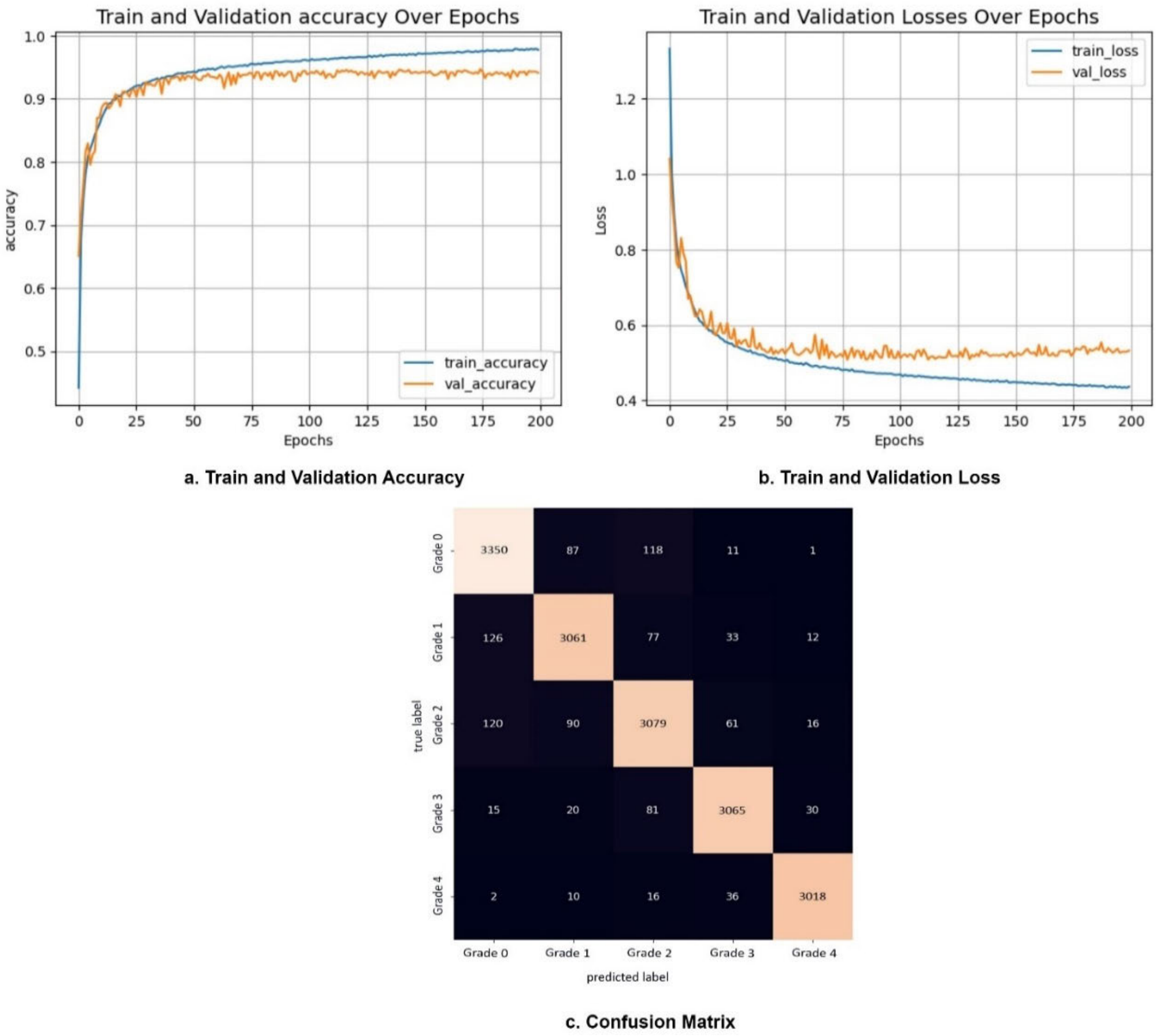
TABLE 14. Transformer model's performance comparison based on accuracy and epoch time.

Transformer Model	Accuracy	Epoch × Time
Involutional neural network	83.58%	200 x 295s
Swin transformer	84.21%	200 x 396s
Vision Transformer	84.81%	200 x 402s
CCT	86.54%	200 x 225s

D. RESULT OF THE OPTIMAL MODEL

1) EVALUATING THE PROPOSED MODEL'S PERFORMANCE

Our suggested CCT model achieved notably better classification accuracy after completing an ablation study on



a. Train and Validation Accuracy

b. Train and Validation Loss

c. Confusion Matrix

FIGURE 12. Visualization of (A) Accuracy curve (B) Loss curve (C) Confusion matrix.

TABLE 15. Computation of different matrixes for the proposed model's performance assessment.

F1 Score	94.2%
Precision	94.4%
Recall	94.2%
Specificity	98.53%
FPR	0.01464
FNR	0.05738
FDR	0.05718
NPV	98.54%
MCC	92.81%

TABLE 16. Individual accuracy for five grades.

Grade	Accuracy
Grade 0	97.10%
Grade 1	97.25%
Grade 2	96.50%
Grade 3	98.26%
Grade 4	99.26%

the base model. Table 15 displays the results of numerous performance metrics, including statistical evaluation for the suggested CCT model. The model got an F1-score of 94.2%, an accuracy of 94.4%, a recall of 94.2%, a specificity of 98.53%, and an accuracy of 94.4% when the test data set was used to test the proposed CTT model. The corresponding values for the FPR, FNR, FDR, NPV, and MCC were 0.01464, 0.05738, 0.05718, 98.54%, and 92.81%. The results of the performance measures show that our proposed model is capable of accurately classifying X-ray images.

TABLE 17. Accuracy comparison with different augmentation techniques.

Description of the original and Photometric augmented datasets and their Accuracy				
Class	Original	Augmented	Total image	Accuracy (%)
Grade 0- Normal	3782	15128	40,928	88.33
Grade 1-Doubtful	2056	8224		
Grade 2-Mild	2441	9764		
Grade 3-Moderate	1408	5632		
Grade 4-Severe	545	2180		
Description of the original and Geometric augmented datasets and their Accuracy				
Class	Original	Augmented	Total image	Accuracy (%)
Grade 0- Normal	3782	15128	40,928	76.56
Grade 1-Doubtful	2056	8224		
Grade 2-Mild	2441	9764		
Grade 3-Moderate	1408	5632		
Grade 4-Severe	545	2180		
Description of the original and Elastic deformation augmented datasets and their Accuracy				
Class	Original	Augmented	Total image	Accuracy (%)
Grade 0- Normal	3782	15,000	75,000	80.22
Grade 1-Doubtful	2056	15,000		
Grade 2-Mild	2441	15,000		
Grade 3-Moderate	1408	15,000		
Grade 4-Severe	545	15,000		
Description of the original and DCGAN augmented datasets and their Accuracy				
Class	Original	Augmented	Total image	Accuracy (%)
Grade 0- Normal	3782	23782	1,10,232	94.58
Grade 1-Doubtful	2056	22056		
Grade 2-Mild	2441	22441		
Grade 3-Moderate	1408	21408		
Grade 4-Severe	545	20545		

Table 16 presents the individual accuracy for all five grades. From Table 16, we can see that grade 0 achieved 97.10% accuracy. Grade 1 and grade 2 obtained an accuracy of 97.25% and 96.50%, respectively. Grade 3 accuracy is considered the second-best with an accuracy of 98.26%. Grade 4 recorded the highest accuracy of 99.29%.

E. A BRIEF SUMMARY OF THE PROCESS OF CREATING IMPROVED DATASETS USING VARIOUS AUGMENTATION METHODS

Table 17 illustrates the use of four distinct augmentation techniques to evaluate the model's performance. We supplement the dataset's original images with elastic deformation [57] approaches, four geometric techniques (vertical flipping, horizontal flipping, rotation 90°, and rotation −90°) [41], four photometric techniques (increasing brightness, reducing brightness, increasing contrast, and reducing contrast) [41], and DCGAN. In order to balance the quantity of images in each class, we immediately applied several augmentation strategies to the dataset. We created a total of 40,928 images using photometric augmentation techniques. We evaluated our proposed model and found the accuracy of 88.33%. We processed 40928 images using geometric approaches and found 76.56% accuracy. After employing the elastic deformation approach to balance the dataset with 75,000 images, we achieved an accuracy of 80.22%. In conclusion, we evaluated DCGAN augmentation methods and generated a total of 1,102,232 images. The DCGAN-generated dataset performs better in our proposed model, with 94.58% accuracy.

F. PERFORMANCE WITH IMAGE REDUCTION

In this study, we checked the reliability of our proposed KOA-CCTNet model by reducing the number of images. We kept reducing the number of images it looked at by 25% in every step to see if it still worked well. We did this test three times with different random sets of images each time to make sure our results were accurate. Figure 11 shows the average accuracy we got from these tests and helping us to understand

TABLE 18. Accuracy comparison with existing literature.

Paper	Name of the dataset	Data Augmentation	Apply image pre-processing techniques	Models	Classification Types	Accuracy
A. S. Mohammed et al. [18]	1.Osteoarthritis Initiative (OAI) 2. Kaggle (Total image = 9786)	N/A	N/A	VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNet V2, and DenseNet121	1. Healthy 2. Doubtful 3. Minimal 4. Moderate 5. Severe	Dataset I - 69% (ResNet101), Dataset II - 83% (ResNet101), Dataset III - 89% (ResNet101)
S. Olsson et al. [22]	Osteoarthritis Initiative (OAI) (Total image = 6403)	N/A	Yes	Convolutional neural network (CNN)	1. Normal 2. Doubtful 3. Severe	87% (CNN)
U. Yunus et al. [24]	Mendeley VI (Total image = 3,795)	N/A	N/A	Open exchange neural network (ONNX) and YOLOv2	1. Normal 2. Doubtful 3. Mild 4. Moderate 5. Severe	90.6% (Their proposed model)
The current study	**1. Kaggle** (400 image) **2. Mendeley I** (8260 image) **3. Mendeley II** (1650 image) **4. AIDA Datahub** (1121 image) **Marge Dataset** (total image 11,431. After removing damage images, the final marge dataset contains 10,232 raw images)	**DCGAN** (generated 1,10,232 augmented images from 10,232 raw images) **Total Dataset: 1,10,232**	**1. LAB Channel.** **2. Adaptive Histogram Equalization (AHE).** **3. Fast Non-Local Means (FNLM)** **4. Resize**	**Proposed Model (KOA-CCTNet)**	**1. Normal** **2. Doubtful** **3. Mild** **4. Moderate** **5. Severe**	**94.58%** (for marge dataset of all five grade) **97.10%** (Grade-0), **97.25%** (Grade-1), **96.50%** (Grade-2), **98.26%** (Grade-3), **99.26%** (Grade-4)

how our model performs with less data. We experimented with 100%, 75%, 50% and 25% images of testing dataset and achieved 94.58%, 92.26%, 90.15%, and 87.55% accuracy. Figure 11 delineate the model accuracy for three times test cases.

G. VISUALIZATION OF CONFUSION MATRIX, ACCURACY CURVE AND LOSS CURVE

Figure 12 illustrates the performance of our proposed model through its accuracy and loss curves. In Figure 12(a), we can see that the training and validation curves come together nicely, showing no substantial gaps between them; this is a good sign indicating no overfitting during the training stage. This positive trend is mirrored in Figure 12(b), where the loss curves also come together nicely, reassuring us that the training did not suffer from overfitting or underfitting issues. Finally, Figure 12(c) visually represents this data through the model's confusion matrices

H. COMPARISON WITH EXISTING WORK

Table 18 provides an overview of the comparison between our proposed model and the existing related works based on accuracy and datasets.

Mohammed et al. [18] introduced four multiclass classification, using three datasets where dataset III produced the best accuracy of 89%. For dataset I and dataset II, their model achieved an accuracy of 69% and 83%, respectively. However, in their work there is a lack of augmentation and image pre-processing techniques. S. Olsson and his team [22], used Osteoarthritis Initiative (OAI) dataset and their model CNN achieved 87% of accuracy. In their work, there is also a lack of augmentation techniques. Other researcher [24] used Mendeley VI dataset and their propose model reach the accuracy of 90.6%. In light of similar earlier research, it may be concluded that multiclass classification using a large dataset and applying the advanced image pre-processing augmentation technique (DCGAN) is more challenging.

While our proposed approach acquired the satisfactory result in with a large datahub in terms of KOA multi-class classification.

V. CONCLUSION

In our study, we aimed to enhance the classification of KOA utilizing X-ray images using a large datahub, leading to the development of the KOA-CCTNet framework. This framework is an innovation based on a modified transformer model, specifically the CCT model, which we selected as our foundational architecture after evaluating four transformer and five transfer learning models. We curated a diverse dataset comprising four distinct sets of X-ray images, each varying in quality and resolution. These images presented challenges due to noise and artifacts, prompting us to employ image processing techniques to improve the quality of all 110,232 images in our dataset. We prioritized creating a balanced dataset, employing augmentation strategies to expand its size and enhance the model's training efficiency. This approach was crucial in addressing the diverse nature of our image dataset and ensuring comprehensive training. Our extensive evaluation included nine different ablation studies, enabling us to optimize the KOA-CCTNet model and address challenges related to training time and dataset size. Remarkably, the KOA-CCTNet demonstrated exceptional performance, achieving a 94.58% accuracy rate. This high level of accuracy was maintained even when collaborating with a reduced number of images, showcasing the model's robustness and reliability. Our study makes significant contributions to the field, including the development of a dependable dataset through unique augmentation strategies and the enhancement of image quality using various pre-processing techniques. Moreover, we provided detailed comparisons between different transformer and transfer learning models, ultimately optimizing the KOA-CCTNet to deliver outstanding results.

VI. LIMITATIONS AND FUTURE SCOPES

The transformer model we introduced in the current study, KOA-CCTNet, demonstrated superior performance in comparison to traditional deep learning models, especially in the multiclass classification of numerous low-pixel X-ray images. Despite its commendable achievements, the KOA-CCTNet model is not without its limitations, presenting opportunities for further enhancements in future studies. Evaluating the model's effectiveness in real-time data scenarios would provide valuable insights into its practical applicability and performance under different conditions. In addition, it is also important to highlight that the KOA-CCTNet model exhibits exceptional capabilities in most testing scenarios, consistently delivering accurate classification results across the five distinct categories of KOA X-ray images. One more potential area for exploring is segmenting the region of interest (ROI) from images, exploring 3D reconstruction, and graphical fields, for example: graph neural networks (GNN), geometric deep learning (GDL), etc. This enhancement could contribute to an even more robust and reliable classification performance. Despite its few limitations, the model stands out for its robustness and reliability, showcasing its potential as a valuable tool in the medical imaging domain.

REFERENCES

[1] A. Cui, H. Li, D. Wang, J. Zhong, Y. Chen, and H. Lu, "Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies," *EClinicalMedicine*, vols. 29–30, Dec. 2020, Art. no. 100587, doi: 10.1016/j.eclinm.2020.100587.

[2] P. P. F. M. Kuijer, H. F. van der Molen, and S. Visser, "A health-impact assessment of an ergonomic measure to reduce the risk of work-related lower back pain, lumbosacral radicular syndrome and knee osteoarthritis among floor layers in The Netherlands," *Int. J. Environ. Res. Public Health*, vol. 20, no. 5, p. 4672, Mar. 2023, doi: 10.3390/ijerph20054672.

[3] V. K. V, V. Kalpana, and G. H. Kumar, "Evaluating the efficacy of deep learning models for knee osteoarthritis prediction based on kellgren-lawrence grading system," *e-Prime - Adv. Electr. Eng., Electron. Energy*, vol. 5, Sep. 2023, Art. no. 100266, doi: 10.1016/j.prime.2023.100266.

[4] S. Castagno, M. Birch, M. van der Schaar, and A. McCaskie. (2024). *Prediction of the Rapid Progression of Knee Osteoarthritis Using Automated Machine Learning: A Novel Precision Health Approach for Chronic Degenerative Diseases*. SSRN. Accessed: Aug. 2, 2024. [Online]. Available: https://ssrn.com/abstract=4561796

[5] A. A. S. Afroze, R. Tamilselvi, and M. G. P. Beham, "Machine learning based osteoarthritis detection methods in different imaging modalities: A review," *Current Med. Imag. Formerly Current Med. Imag. Rev.*, vol. 19, no. 14, pp. 1628–1642, Dec. 2023, doi: 10.2174/1573405619666230130143020.

[6] Y. C. Park, K. J. Park, B. H. Goo, J. H. Kim, B. K. Seo, and Y. H. Baek, "Oriental medicine as collaborating treatments with conventional treatments for knee osteoarthritis: A PRISMA-compliant systematic review and meta-analysis," *Medicine*, vol. 102, no. 29, 2023, Art. no. E34212, doi: 10.1097/MD.0000000000034212.

[7] S. V Chaugule, S. V Chaugule, and V. S. Malemath, "Towards identifying key features in the classification of knee osteoarthritis—An enhanced feature fusion based deep network model," Tech. Rep., 2023.

[8] Y. Soda, T. Kano, and M. Nakamura, "Kinematically aligned total knee arthroplasty for valgus osteoarthritis of more than 10° is it still a 'challenging surgery?'" *Open J. Orthopedics*, vol. 13, no. 9, pp. 355–369, 2023, doi: 10.4236/ojo.2023.139035.

[9] H. Khalid, M. Hussain, M. A. Al Ghamdi, T. Khalid, K. Khalid, M. A. Khan, K. Fatima, K. Masood, S. H. Almotiri, M. S. Farooq, and A. Ahmed, "A comparative systematic literature review on knee bone reports from MRI, X-rays and CT scans using deep learning and machine learning methodologies," *Diagnostics*, vol. 10, no. 8, p. 518, Jul. 2020, doi: 10.3390/diagnostics10080518.

[10] J. J. Bjerre-Bastos, M. A. Karsdal, M. Boesen, H. Bliddal, A. Bay-Jensen, J. R. Andersen, and A. R. Bihlet, "The acute and long-term impact of physical activity on biochemical markers and MRI measures in osteoarthritis—Perspectives for clinical osteoarthritis research," *Transl. Sports Med.*, vol. 3, no. 5, pp. 384–394, Sep. 2020, doi: 10.1002/tsm2.155.

[11] Y. X. Teoh, A. Othmani, S. L. Goh, J. Usman, and K. W. Lai, "Deciphering knee osteoarthritis diagnostic features with explainable artificial intelligence: A systematic review," 2023, *arXiv:2308.09380*.

[12] V. K. V, V. Kalpana, and G. H. Kumar, "Evaluating the efficacy of deep learning models for knee osteoarthritis prediction based on kellgren-lawrence grading system," *e-Prime - Adv. Electr. Eng., Electron. Energy*, vol. 5, Sep. 2023, Art. no. 100266.

[13] Y. Nasser, M. El Hassouni, D. Hans, and R. Jennane, "A discriminative shape-texture convolutional neural network for early diagnosis of knee osteoarthritis from X-ray images," *Phys. Eng. Sci. Med.*, vol. 46, no. 2, pp. 827–837, Jun. 2023, doi: 10.1007/s13246-023-01256-1.

[14] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2071–2081, doi: 10.1609/aaai.v36i2.20103.

[15] K. Mohiuddin, M. A. Alam, M. M. Alam, P. Welke, M. Martin, J. Lehmann, and S. Vahdati, "Retention is all you need," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2023, pp. 4752–4758, doi: 10.1145/3583780.3615497.

[16] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[17] G. K. M and A. D. Goswami, "Automatic classification of the severity of knee osteoarthritis using enhanced image sharpening and CNN," *Appl. Sci.*, vol. 13, no. 3, p. 1658, Jan. 2023, doi: 10.3390/app13031658.

[18] A. S. Mohammed, A. A. Hasanaath, G. Latif, and A. Bashar, "Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed X-ray images," *Diagnostics*, vol. 13, no. 8, p. 1380, Apr. 2023, doi: 10.3390/diagnostics13081380.

[19] S. S. Abdullah and M. P. Rajasekaran, "Automatic detection and classification of knee osteoarthritis using deep learning approach," *La radiologia medica*, vol. 127, no. 4, pp. 398–406, Apr. 2022, doi: 10.1007/s11547-022-01476-7.

[20] S. A. El-Ghany, M. Elmogy, and A. A. A. El-Aziz, "A fully automatic fine tuned deep learning model for knee osteoarthritis detection and progression analysis," *Egyptian Informat. J.*, vol. 24, no. 2, pp. 229–240, Jul. 2023, doi: 10.1016/j.eij.2023.03.005.

[21] T. Tariq, Z. Suhail, and Z. Nawaz, "Knee osteoarthritis detection and classification using X-rays," *IEEE Access*, vol. 11, pp. 48292–48303, 2023, doi: 10.1109/ACCESS.2023.3276810.

[22] S. Olsson, E. Akbarian, A. Lind, A. S. Razavian, and M. Gordon, "Automating classification of osteoarthritis according to kellgren-lawrence in the knee using deep learning in an unfiltered adult population," *BMC Musculoskeletal Disorders*, vol. 22, no. 1, pp. 1–8, Dec. 2021, doi: 10.1186/s12891-021-04722-7.

[23] A. Qadir, R. Mahum, and S. Aladhadh, "A robust approach for detection and classification of KOA based on BILSTM network," *Comput. Syst. Sci. Eng.*, vol. 47, no. 2, pp. 1365–1384, 2023, doi: 10.32604/csse.2023.037033.

[24] U. Yunus, J. Amin, M. Sharif, M. Yasmin, S. Kadry, and S. Krishnamoorthy, "Recognition of knee osteoarthritis (KOA) using YOLOv2 and classification based on convolutional neural network," *Life*, vol. 12, no. 8, p. 1126, 2022.

[25] P. Chen, "Knee osteoarthritis severity grading dataset," Mendeley, Sep. 2018, doi: 10.17632/56rmx5bjcr.1.

[26] S. Gornale and P. Patravali, "Digital knee X-ray images," Mendeley, Jun. 2020, doi: 10.17632/t9ndx37v5h.1.

[27] kaggle. *CGMH Osteoarthritis Images*. Accessed: Jul. 31, 2020. [Online]. Available: https://www.kaggle.com/datasets/tommyngx/cgmh-oa

[28] M. Gordon and E. Akbarian, "Knee osteoarthritis classification according to Kellgren-Lawrence," Tech. Rep., 2021, doi: 10.23698/aida/koa2021.

[29] S. S. Gornale, P. U. Patravali, and R. R. Manza, "Detection of osteoarthritis using knee X-ray image analyses: A machine vision based approach," *Int. J. Comput. Appl.*, vol. 145, no. 1, pp. 20–26, Jul. 2016, doi: 10.5120/ijca2016910544.

[30] W. Fang, F. Zhang, V. S. Sheng, and Y. Ding, "A method for improving CNN-based image recognition using DCGAN," *Comput., Mater. Continua*, vol. 57, no. 1, pp. 167–178, 2018, doi: 10.32604/cmc.2018.02356.

[31] M. Puttagunta and R. Subban, "A novel COVID-19 detection model based on DCGAN and deep transfer learning," *Proc. Comput. Sci.*, vol. 204, pp. 65–72, Jan. 2022, doi: 10.1016/j.procs.2022.08.008.

[32] D. J. Bora, A. K. Gupta, and F. A. Khan, "Comparing the performance of L*A*B* and HSV color spaces with respect to color image segmentation," 2015, *arXiv:1506.01472*.

[33] A. V. Reddy, R. Thiruvengatanadhan, M. Srinivas, and P. Dhanalakshmi, "Region based segmentation with enhanced adaptive histogram equalization model with definite feature set for sugarcane leaf disease classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 1s, pp. 428–442, 2024.

[34] P. Yugander, K. Abhishek, P. S. Reddy, G. Manideep, T. Sahithi, and M. Jagannath, "Extraction of blood vessels from retinal fundus images using maximum principal curvatures and adaptive histogram equalization," in *Proc. 1st Int. Conf. Electr. Electron. Inf. Commun. Technol. (ICEEICT)*, Feb. 2022, doi: 10.1109/ICEEICT53079.2022.9768517.

[35] Y. Mousania, S. Karimi, and A. Farmani, "Optical remote sensing, brightness preserving and contrast enhancement of medical images using histogram equalization with minimum cross-entropy-Otsu algorithm," *Opt. Quantum Electron.*, vol. 55, no. 2, pp. 1–22, Feb. 2023, doi: 10.1007/s11082-022-04341-z.

[36] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 60–65.

[37] A. Qi, D. Zhao, F. Yu, A. A. Heidari, Z. Wu, Z. Cai, F. Alenezi, R. F. Mansour, H. Chen, and M. Chen, "Directional mutation and crossover boosted ant colony optimization with application to COVID-19 X-ray image segmentation," *Comput. Biol. Med.*, vol. 148, Sep. 2022, Art. no. 105810, doi: 10.1016/j.compbiomed.2022.105810.

[38] S. Montaha, S. Azam, A. K. M. R. H. Rafid, P. Ghosh, M. Z. Hasan, M. Jonkman, and F. De Boer, "BreastNet18: A high accuracy fine-tuned VGG16 model evaluated using ablation study for diagnosing breast cancer from enhanced mammography images," *Biology*, vol. 10, no. 12, p. 1347, Dec. 2021.

[39] K. Fatema, S. Montaha, M. A. H. Rony, S. Azam, M. Z. Hasan, and M. Jonkman, "A robust framework combining image processing and deep learning hybrid model to classify cardiovascular diseases using a limited number of paper-based complex ECG images," *Biomedicines*, vol. 10, no. 11, p. 2835, Nov. 2022, doi: 10.3390/biomedicines10112835.

[40] H. A. Sanghvi, R. H. Patel, A. Agarwal, S. Gupta, V. Sawhney, and A. S. Pandya, "A deep learning approach for classification of COVID and pneumonia using DenseNet-201," *Int. J. Imag. Syst. Technol.*, vol. 33, no. 1, pp. 18–38, Jan. 2023, doi: 10.1002/ima.22812.

[41] S. Montaha, S. Azam, A. K. M. R. H. Rafid, M. Z. Hasan, A. Karim, K. M. Hasib, S. K. Patel, M. Jonkman, and Z. I. Mannan, "MNet-10: A robust shallow convolutional neural network model performing ablation study on medical images assessing the effectiveness of applying optimal data augmentation technique," *Frontiers Med.*, vol. 9, Aug. 2022, Art. no. 924979.

[42] F. J. M. Shamrat, S. Azam, A. Karim, K. Ahmed, F. M. Bui, and F. De Boer, "High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106646, doi: 10.1016/j.compbiomed.2023.106646.

[43] R. Indraswari, R. Rokhana, and W. Herulambang, "Melanoma image classification based on MobileNetV2 network," *Proc. Comput. Sci.*, vol. 197, pp. 198–207, Jan. 2022, doi: 10.1016/j.procs.2021.12.132.

[44] Y. E. Almalki, M. Zaffar, M. Irfan, M. A. Abbas, M. Khalid, K. S. Quraishi, T. Ali, F. Alshehri, S. K. Alduraibi, A. A. Asiri, M. A. A. Basha, A. Alduraibi, M. K. Saeed, and S. Rahman, "A novel-based Swin transfer based diagnosis of COVID-19 patients," *Intell. Autom. Soft Comput.*, vol. 35, no. 1, pp. 163–180, 2023, doi: 10.32604/iasc.2023.025580.

[45] J. Feng, H. Zhang, M. Geng, H. Chen, K. Jia, Z. Sun, Z. Li, X. Cao, and B. W. Pogue, "X-ray cherenkov-luminescence tomography reconstruction with a three-component deep learning algorithm: Swin transformer, convolutional neural network, and locality module," *J. Biomed. Opt.*, vol. 28, no. 2, pp. 1–18, Feb. 2023, doi: 10.1117/1.jbo.28.2.026004.

[46] A. Marefat, M. Marefat, J. H. Joloudari, M. A. Nematollahi, and R. Lashgari, "CCTCOVID: COVID-19 detection from chest X-ray images using compact convolutional transformers," *Frontiers Public Health*, vol. 11, Feb. 2023, Art. no. 1025746, doi: 10.3389/fpubh.2023.1025746.

[47] I. U. Khan, S. Azam, S. Montaha, A. A. Mahmud, A. K. M. R. H. Rafid, M. Z. Hasan, and M. Jonkman, "An effective approach to address processing time and computational complexity employing modified CCT for lung disease classification," *Intell. Syst. with Appl.*, vol. 16, Nov. 2022, Art. no. 200147, doi: 10.1016/j.iswa.2022.200147.

[48] C.-M. Lo and K.-L. Lai, "Deep learning-based assessment of knee septic arthritis using transformer features in sonographic modalities," *Comput. Methods Programs Biomed.*, vol. 237, Jul. 2023, Art. no. 107575.

[49] O. Uparkar, J. Bharti, R. K. Pateriya, R. K. Gupta, and A. Sharma, "Vision transformer outperforms deep convolutional neural network-based model in classifying X-ray images," *Proc. Comput. Sci.*, vol. 218, pp. 2338–2349, Jan. 2023, doi: 10.1016/j.procs.2023.01.209.

[50] M. M. Al Rahhal, Y. Bazi, R. M. Jomaa, A. AlShibli, N. Alajlan, M. L. Mekhalfi, and F. Melgani, "COVID-19 detection in CT/X-ray imagery using vision transformers," *J. Personalized Med.*, vol. 12, no. 2, p. 310, Feb. 2022, doi: 10.3390/jpm12020310.

[51] J. J. Mondal, M. F. Islam, S. Zabeen, and M. A. Manab, "InvoPotNet: Detecting pothole from images through leveraging lightweight involutional neural network," in *Proc. 25th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Cox's Bazar, Bangladesh, 2022, pp. 599–604, doi: 10.1109/ICCIT57492.2022.10055818.

[52] W. Ou and S.-I. Kamata, "Skin lesion classification based on involution neural networks with Triplet++ attention generator," in *Proc. 8th Int. Conf. Biomed. Signal Image Process.*, Jul. 2023, pp. 1–6.

[53] I. U. Khan, M. A. K. Raiaan, K. Fatema, S. Azam, R. U. Rashid, S. H. Mukta, M. Jonkman, and F. De Boer, "A computer-aided diagnostic system to identify diabetic retinopathy, utilizing a modified compact convolutional transformer and low-resolution images to reduce computation time," *Biomedicines*, vol. 11, no. 6, p. 1566, May 2023, doi: 10.3390/biomedicines11061566.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[55] M. A. K. Raiaan, K. Fatema, I. U. Khan, S. Azam, M. R. U. Rashid, M. S. H. Mukta, M. Jonkman, and F. De Boer, "A lightweight robust deep learning model gained high accuracy in classifying a wide range of diabetic retinopathy images," *IEEE Access*, vol. 11, pp. 42361–42388, 2023, doi: 10.1109/ACCESS.2023.3272228.

[56] K. Fatema, M. A. H. Rony, S. Azam, M. S. H. Mukta, A. Karim, M. Z. Hasan, and M. Jonkman, "Development of an automated optimal distance feature-based decision system for diagnosing knee osteoarthritis using segmented X-ray images," *Heliyon*, vol. 9, no. 11, Nov. 2023, Art. no. e21703.

[57] S. Rana, M. J. Hosen, T. J. Tonni, M. A. H. Rony, K. Fatema, M. Z. Hasan, M. T. Rahman, R. T. Khan, T. Jan, and M. Whaiduzzaman, "DeepChestGNN: A comprehensive framework for enhanced lung disease identification through advanced graphical deep features," *Sensors*, vol. 24, no. 9, p. 2830, Apr. 2024, doi: 10.3390/s24092830.

MUSHRAT JAHAN received the bachelor's degree in computer science and engineering from Daffodil International University, Dhaka, Bangladesh. She is actively involved in research activities across various domains, especially in health informatics, computer vision, machine learning, deep learning, and artificial intelligence. She has published several research papers in international conferences.

MD. ZAHID HASAN (Member, IEEE) received the M.Sc. degree in information and communication engineering from the University of Rajshahi and the Ph.D. degree in computer science and engineering from Jahangirnagar University. He is currently an Associate Professor with the Department of Computer Science and Engineering, Daffodil International University, Bangladesh. His research interests include computer vision, health informatics, machine learning, artificial intelligence, and decision theory.

ISMOT JAHAN SAMIA received the bachelor's degree in computer science and engineering from Daffodil International University, Dhaka, Bangladesh. She is now focused on research in several dynamic fields, including health informatics, computer vision, as well as machine and deep learning, contributing to innovations in artificial intelligence.

KANIZ FATEMA received the bachelor's degree in computer science and engineering from Daffodil International University, Dhaka, Bangladesh. She is currently serving as a Research coordinator at Health Informatics Research Laboratory and also working as a Research Assistant (RA) at Charles Darwin University. She is actively involved in research activities, especially in health informatics, computer vision, machine learning, deep learning, and artificial intelligence-based systems. She has published several research papers in journals (Scopus) and international conferences.

MD. AWLAD HOSSEN RONY received the bachelor's degree from the Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. He is currently serving as a Research coordinator at Health Informatics Research Laboratory and also working as a Research Assistant at Charles Darwin University, Australia, where he is deeply involved in cutting-edge research across various domains, including computer vision, health informatics, machine learning, deep learning, image processing, and artificial intelligence.

MOHAMMAD SHAMSUL AREFIN (Senior Member, IEEE) received the Doctor of Engineering degree in information engineering from Hiroshima University, Japan, with the support of the Scholarship of MEXT, Japan. As a part of his Ph.D. research, he was with the IBM Yamato Software Laboratory, Japan. He was the Head of the Department. He is currently with the Department of Computer Science and Engineering (CSE), Chittagong University of Engineering and Technology, Bangladesh. He visited Japan, Indonesia, Malaysia, Bhutan, Singapore, South Korea, Egypt, India, Saudi Arabia, and China for different professional and social activities at United International University. He has published several research papers in journals (Scopus) and international conferences. He has more than 110 refereed publications in international journals, book series, and conference proceedings. His research interests include privacy-preserving data publishing and mining, distributed and cloud computing, big data management, multilingual data management, semantic web, object-oriented systems development, and IT for agriculture and the environment. He is a member of ACM and a fellow of IEB and BCS. He is the Organizing Chair of BIM 2021, the TPC Chair of ECCE 2017, the Organizing Co-Chair of ECCE 2019, and the Organizing Chair of BDML 2020.

AHMED MOUSTAFA has worked in Sydney, USA, and Cairo, before moving to Bond University. He has obtained grant funding from Australia, the USA, Qatar, the United Arab Emirates, Türkiye, and other countries. He publishes collaboratively with 71 colleagues and has more than 510 co-authors, from 35 institutions in 14 countries. He has published over 240 articles in high-ranking journals, including *Science*, PNAS, *The Journal of Neuroscience*, *Brain*, *Neuroscience and Biobehavioral Reviews*, *Parkinson's Disease* (Nature), and *Neuron*, among others. His research interests include mental health, machine learning, and bioinformatics, including clinical disorders, such as drug addiction and Alzheimer's disease.