Noname manuscript No.
(will be inserted by the editor)

Grading Knee Osteoarthritis: Modernizing Treatment with Transfer Learning

Anshul Kumar · Bhavik Niranjane · Rahul Bhure ·

Received: date / Accepted: date

Abstract Osteoarthritis, a prevalent degenerative condition, greatly affects the elderly population, particularly in the context of knee osteoarthritis (KOA). Traditional X-ray analysis using the Kellgren and Lawrence grading system presents challenges in accurately diagnosing KOA due to its subjective nature. This study introduces an innovative approach leveraging artificial intelligence, specifically deep learning techniques, to address these challenges. By employing Convolutional Neural Networks (CNNs), transfer learning, and Ensemble Techniques, this study aims to improve the accuracy of KOA detection and classification based on the Knee Osteoarthritis Dataset with severity grading, which consists of X-ray images of knee joints classified into 5 classes on the Kellgren-Lawrence grading system. This method facilitates early and precise diagnosis, enabling timely interventions that may slow disease progression. Comparative analyses with previous models demonstrate the effectiveness of the proposed approach. This research has achieved 84% accuracy for binary classification, 90% for tertiary classification, and 71% for 5-category classi-

Anshul Kumar
Indian Institute of Information Technology, Pune, India.
E-mail: 112215024@cse.iiitp.ac.in

Bhavik Niranjane
Indian Institute of Information Technology, Pune, India.
E-mail: 112215042@cse.iiitp.ac.in

Rahul Bhure
Indian Institute of Information Technology, Pune, India.
E-mail: 112215036@cse.iiitp.ac.in

fication. Incorporating deep learning in KOA diagnosis enhances diagnostic accuracy and reduces dependence on manual expertise, promising significant advancements in medical imaging and patient care in osteoarthritis management. This paper also presents the comparative study of proposed methods with existing methods to show the effectiveness of the work.

Keywords Knee Osteoarthritis · Deep Learning · Convolutional Neural Networks · Medical Imaging

1 Introduction

Osteoarthritis, commonly known as wear-and-tear arthritis, is a degenerative disease marked by the progressive erosion of cartilage, the natural cushioning in joints. This erosion weakens the relationship between bones in the joint, leading to pain, swelling, stiffness, and reduced mobility. In some cases, friction from the lack of cushioning can cause bone spurs, worsening symptoms. Primarily affecting the elderly, osteoarthritis occurs in two types: primary (with no clear cause) and secondary (from injury or disease). Risk factors include age, weight, genetics, gender, repetitive strain injuries, and medical conditions like rheumatoid arthritis and metabolic syndrome.

Diagnosis often involves imaging techniques such as X-rays and MRI scans, which help assess bone and cartilage damage and detect bone spurs. Early diagnosis is crucial for effective intervention, which may include pain management, lifestyle adjustments, exercise, and, in some cases, surgery.

Deep learning presents a promising tool for knee osteoarthritis (KOA) detection, analyzing large sets of X-rays to identify KL grade patterns with potential for high accuracy. This approach could enhance radiologist workflows by highlighting key features in X-rays. Using the Kellgren and Lawrence (KL) grading system, deep learning models mitigate the subjectivity and variability of traditional X-ray interpretation.

This study explores deep learning for KOA detection through KL grading. Data augmentation techniques for knee X-rays address data size and variability limitations, while transfer learning and ensemble methods enhance model robustness. The proposed model classifies OA symptoms into KL grading categories, distinguishing between healthy and diseased states to facilitate early diagnosis and management.// The key contributions of this paper are:

2 Literature Review

Computer vision technology has emerged as a robust tool for diagnosing knee osteoarthritis (OA) via X-ray image classification. Analogous to its application in facial recognition or object detection, computer vision scrutinizes knee X-rays to discern OA indicators and ascertain severity levels. Leveraging an extensive corpus of medical literature encompassing scholarly works and journal publications, researchers acquire foundational insights into the historical

trajectory, contemporary understanding, and persisting challenges inherent in OA diagnosis. This corpus serves as a bedrock for the development and refinement of computer vision algorithms tailored for X-ray analysis, thereby augmenting efficiency and precision in knee OA classification endeavors. The integration of image processing techniques into X-ray interpretation harbors significant potential for automating facets of OA classification, heralding expedited diagnoses, refined disease progression monitoring, and elevated standards of patient care.

Cheng-Tzu Wang et al. have proposed an osteoarthritis classification deep-learning model employing YOLOv4 and Conv-BN-ReLU block based model. Their research aimed to swiftly and accurately classify knee osteoarthritis with the help of osteoarthritis initiative (OAI) data set comprising of 8964 knees and clinical AP radiographs of 246 knees from FEMH with 5 KL grade classes, achieving an impressive 78% accuracy.

Paulione S. Q. Yeoh et al. research works has proposed 3D assisted knee osteoarthritis classification leveraging the CNN models(ResNet, DenseNet, VGG and AlexNet). The Osteoarthritis Initiative (OAI) dataset is used, which contains 4,796 images. ResNet18 achieved the highest AUC score of 94.5%.

Haresh Rajamohan et al. work proposed the prediction of total knee replacement with knee MRIs using a CNN architecture. This research is based on the OAI dataset classified into IW-TSE, FS-IW-TSE, and DESS, having 4796 images in total, leveraging a 90% AUC score.

Nide Nasir et al. proposed multi-modal image classification of COVID-19 images captured using computed tomography and X-ray scans using transfer learning models such as VGG16, ResNet50, InceptionResNetV2 and MobileNetV2. The research is based on the covid-xray-dataset, which binary classifies into "Covid" and "Non-Covid". The accuracy turned out to be 97.8%.

Aleksei et al. works proposed Knee Osteoarthritis Diagnosis from Plain Radiographs mounted on deep learning. This research is based on a 5960 image-oriented database, the Osteoarthritis Initiative dataset, using Deep Siamese CNN architecture. The research claimed a quadratic Kappa coefficient of 0.83 and an average multiclass accuracy of 66.71%.

Hua Wang et al. research work proposed ankle fracture detection deploying EfficientNetB5, ResNet50 with the Squeeze-and-Excitation Network (SENet). In this research, CT images of ankle fractures images are collected, numbering 987 , deemed apt for research purposes, segmented into 255 images depicting fractures and 732 illustrating normal ankles. The result for ResNet50 with SENet was 93%, and EfficientNetB5 was 90%.

R. V. Manjunath et al.research work proposed a detection and classification model of liver disease using CT images based on a deep learning model called modified Unet 60. A public dataset 3Dircadb is utilized containing 864 images, 360 images belonging to Metastasis cancer, and 360 images affiliated to Cholangiocarcinoma. This research attained an accuracy of 98.61% and a dice score of 98.59%.

Qingji Guan et al. proposed the Chest X-ray Image Classification with Noisy Labels using Heteroscedastic Modeling Method. This research embraces an ChestX-ray2017 dataset containing 5856 anterior-posterior Chest X-ray images including 14 pathologies: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, PT, and Hernia. This model achieved an average AUC score of 14 pathologies as 81.2%.
Ahmed Khalid et al. works proposed to predict Knee Osteoarthritis Grades based on fusion features FFNN (CNN and handcrafted). This research is based on the OAI and RCU datasets containing 5 classes of severity grading Healthy, Doubtful, Minimal, Moderate, and Severe, having a total of 11,436 images. It gained 99.1% recognition accuracy.

This comprehensive literature review offers a deep dive into the current landscape of deep learning models for knee osteoarthritis (OA) detection and classification. By synthesizing a wide range of research, we gain insights into the strengths of various deep learning architectures for KL grading, allowing us to identify the most promising approaches moving forward. The review also sheds light on the impact of data augmentation techniques in enriching datasets and transfer learning's ability to accelerate training. However, limitations like potential data bias and the "black box" nature of some models are highlighted, pinpointing areas for future research. This review paves the way for developing more robust and trustworthy deep-learning solutions for knee OA diagnosis by identifying these challenges and analyzing promising areas like integrating clinical data or improving explainability.

3 Materials and methods

The strategy we will be using is a simple and efficient one. We will first preprocess the data set using different techniques to maximize the use the capacity of the dataset and avoid overfitting the CNNs model while training them. Then, with the help of transfer learning and ensemble learning train our models on the dataset and then test their performance based on different criteria. Fig. shows the overall flow of the work.

3.1 Dataset Description

The dataset used here is the Knee Osteoarthritis Dataset with severity grading, which consists of X-ray images of knee joints divided into 5 classes. The grading system used in the dataset is the Kellgren-Lawrence grading system, which is a popular technique for determining the degree of osteoarthritis in the knee based on X-ray pictures. In the 1950s, Drs. John Kellgren and Jeffrey Lawrence created it. Based on the existence of specific radiographic characteristics, the method assigns five grades, from 0 to 4, to the severity of knee

osteoarthritis. Grade 0 represents no indications of osteoarthritis. There are no osteophytes (bone spurs) or other anomalies, and the joint space is normal. Grade 1 represents doubtful condition that there may be osteophytic lipping and dubious joint narrowing, which could indicate very early osteoarthritis symptoms. Grade 2 represents minimal osteophytes that are clearly present, indicating the onset of osteoarthritis. Additionally, joint space narrowing could occur. Grade 3 represents a moderate condition in which there are several osteophytes, noticeable narrowing of the joint space, and mild sclerosis (hardening) of the bone. This suggests a mild case of osteoarthritis. Grade 4 represents severe condition that there is severe sclerosis, large osteophytes are apparent, and there is noticeable joint space constriction.

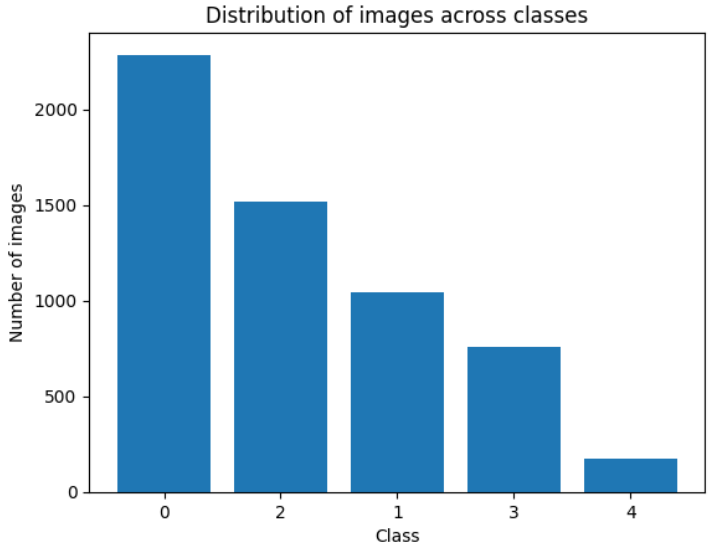

Fig. 1 Training Data Distribution

The dataset consists of four directories: Auto_test, Train, Test, and Validation. Each of these directories consists of X-ray images belonging to five classes. Auto_test directory contains 604 images of Grade0, 275 images of Grade1, 403 images of Grade2, 200 images of Grade3, 44 images of Grade4. Train directory contains 2286 images of Grade0, 1046 images of Grade1, 1516 images of Grade2, 757 images of Grade3, 173 images of Grade4. Validation directory contains 328 images of Grade0, 153 images of Grade1, 212 images of Grade2, 106 images of Grade3, 27 images of Grade4. Test directory contains 639 images of Grade0, 296 images of Grade1, 447 images of Grade2, 223 images of Grade3, 51 images of Grade4.

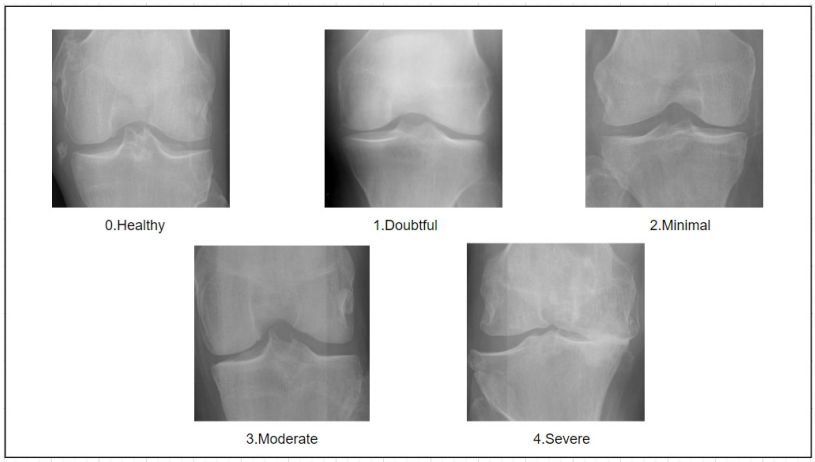

Fig. 2 Dataset Classes

3.2 Data Preprocessing

Data Preprocessing[**?**] is a big part when it comes to training a deep learning model. With the help of data augmentation, we can use data sets to their maximum extent.

3.2.1 Five Class Classification:

Random Erasing, a powerful data augmentation technique, was utilized to enhance the generalization of deep learning models in image classification tasks. This method involves randomly replacing a rectangular region in the input image with random pixel values, simulating occlusions or corruptions. By exposing the models to diverse inputs during training, Random Erasing promotes robust feature learning and improves model performance on unseen data. Tuned hyperparameters optimized the application of Random Erasing, contributing significantly to the models' classification accuracy, especially in challenging tasks like the 5-class classification.

3.2.2 Three Class Classification:

We have performed three class classifications by considering Grade 2, Grade 3, and Grade 4 classes. Random Erasing, a powerful data augmentation technique, was utilized to enhance the generalization of deep learning models in image classification tasks. This method randomly replaces a rectangular region in the input image with random pixel values, simulating occlusions or corruptions. By exposing the models to diverse inputs during training, Random Erasing promotes robust feature learning and improves model performance

on unseen data. Tuned hyperparameters optimized the application of Random Erasing, contributing significantly to the models' classification accuracy, especially in challenging tasks like the 5-class classification.

3.2.3 Binary Classification:

For binary classification, this work have merged Grade0 and Grade1 into class 'Healthy' and Grade2, Grade3, Grade4 into class 'Unfit'. After classifying the whole dataset into two classes, we are trimming these two classes such that there are 500 images in each of them. Using the ImageDataGenerator class and its parameters—horizontal flip, rotation, width shift, height shift, and zoom range—the images for each class are balanced. Once every class that is taken into account has been balanced, Once every class that is taken into account has been balanced, Image Augmentation is applied on training and validation data in order to extract more insightful information from them.

3.3 Transfer Learning

A machine learning technique called transfer learning makes use of information from related activities to enhance a model's performance on a target task. It is especially helpful when the source and target jobs are somewhat related but may differ in some specific areas. Transfer learning reduces the requirement for large amounts of labeled data and training time in the target domain by allowing knowledge from a prior task to be reused instead of starting from scratch when training a new model. Transfer learning in the context of deep learning entails leveraging pre-trained models that have been trained on sizable and varied datasets, such as Convolutional Neural Networks (CNNs), for visual tasks.

3.3.1 InceptionV3

A deep convolutional neural network (CNN) architecture called Inception V3 was created with image identification tasks in mind. It was created by Google researchers and is a noteworthy development in the field of computer vision. The architecture allows it to automatically learn hierarchical features from unprocessed visual input by utilizing many layers of convolutional and pooling procedures. The inventive usage of "Inception modules," which are made up of parallel convolutional layers with various filter sizes, is what sets Inception V3 apart. These modules improve the network's capacity to identify intricate patterns in images by enabling it to record a broad variety of characteristics at different spatial scales.

4 Experiments and Results

This section is dedicated to presenting the results obtained by training and testing the current dataset on using deep-learning architectures.

4.1 Hardware and Software Setup

4.2 Evaluation Criteria

The results are evaluated using various important metrics, such as accuracy and confusion matrix. The definition of all metrics is given here:

Accuracy : It is a fraction of the total correct predictions.[**?**]

$$Accuracy = \frac{(TP+TN)}{(TP+FN)+(FP+TN)} \quad (1)$$

Results were recorded and compared against other models. We have used the Confusion Matrix to evaluate the models tested.

Confusion Matrix : A confusion matrix is an NxN matrix where N represents the number of classes being predicted

5 Results and Discussion

5.1 2-Class Classification

In the 2-class classification task, diverse deep learning models were evaluated to identify the most effective architecture for achieving high accuracy. The models considered included DenseNet121, Xception, ResNet101, InceptionNetV3, InceptionResNetV2, and EfficientNetB5. Each model was trained for 40 epochs with an adaptive learning rate of 0.01 and a batch size of 32.

Training accuracy and loss metrics were monitored for all models. DenseNet121 displayed a moderate increase in training accuracy, though its training loss reduction was less pronounced compared to other models. InceptionNetv3 and ResNet101 exhibited more rapid improvements in training accuracy, but their training loss curves suggested occasional overfitting. Xception and Inception-ResNetV2 both showed stable training processes with consistent accuracy improvements and gradual loss decreases.

Other models, such as DenseNet121 and EfficientNetB5, also demonstrated significant performance improvements. Both models achieved an accuracy of 84%, with their training accuracy and loss curves, reflecting a steady learning progression. The validation accuracy and loss for these models,also indicated good generalization with minimal overfitting.

These results highlight the efficacy of these architectures in handling multi-class classification tasks. The consistent batch size and the fine-tuned learning rates across all models ensured an equitable comparison, leading to these insightful observations on model performance.

5.2 5-Class Classification

In the 5-class classification task, the deep learning models exhibited varying performance levels. DenseNet201 (Figure **??**) achieved an accuracy of 69.63% after 69 epochs of training with a learning rate of 0.0001 and a batch size of 32. This result (Figure **??**) showcases the model's ability to effectively capture the features. Table **??** shows hyperparameters used.

5.3 Discussion

The deep learning models were evaluated on 2-class, 3-class, and 5-class classification tasks. In the 5-class classification task, DenseNet201 achieved an accuracy of 69.63%, outperforming DenseNet121 and InceptionV3. The models' training and validation accuracy graphs and confusion matrices provided insights into their training dynamics and classification performance.
For the 3-class classification task, DenseNet201 and DenseNet121 exhibited high accuracies of 89.5% and 86% respectively. In the 2-class classification task, EfficientNetB5 achieved an accuracy of 84%, outperforming other models. The ensemble of DenseNet201 and DenseNet121 models achieved an accuracy of 70.2% in the 5-class classification task. InceptionResNetV2 performed exceptionally well in the 3-class classification task, achieving an accuracy of 84.74%. VGG16 also demonstrated a high accuracy of 83.35% in the 3-class classification task.
Overall, the results indicate the effectiveness of DenseNet architectures in multi-class classification tasks. The comparison with other popular models highlights the strengths of DenseNet models in achieving high accuracies across different classification challenges. The training and validation accuracy graphs and confusion matrices offer valuable insights into the models' training dynamics and classification performance.

6 Conclusion and future Scope

In this study, we adopted an effective method combining prior knowledge and transfer learning to train deep learning models for knee osteoarthritis classification. Our analysis reveals important facts about various divisions of labor. More specifically, the EfficientNet model achieved 83.94% accuracy in two-class classification, while the DenseNet201 model achieved 89.5% accuracy in three classes. Additionally, the Ensemble model(DenseNet201 and DenseNet121) achieved an accuracy of 70.2% on the challenging five-class task. These findings show the effectiveness of the de9ep learning method in correctly classifying the level of knee osteoarthritis. Simple models such as EfficientNet perform well in binary classification, while more complex models such as DenseNet201 and Ensemble show the best accuracy in multi-class. Training analysis and validation of trends, as well as analysis of confusion matrices, provide insight into the

model and operational nuances. This highlights the importance of considering model selection and hyperparameter optimization to achieve best results in different tasks. Check the possibility of diagnosis. Using the latest technology and rigorous testing, we are able to produce accurate and reliable automatic diagnostic systems in the medical field. Going forward, continued research aimed at improving design methods and discovering new techniques will hold great promise for improving clinical imaging in musculoskeletal therapy.

The future scope of early detection of knee osteoarthritis using transfer learning holds significant potential. Firstly, refining transfer learning techniques could optimize model performance for this specific application. Additionally, integrating multimodal data could enhance the accuracy and reliability of predictions. Personalized risk assessment models tailored to individual patient characteristics offer promise for targeted interventions. Integration into clinical decision support systems could aid healthcare professionals in making informed decisions. Validation studies are crucial for assessing performance and generalizability across diverse clinical settings. Successful validation could pave the way for clinical translation and widespread implementation. Overall, these advancements have the potential to revolutionize knee osteoarthritis management and improve patient outcomes.

With many new algorithms and models being developed every day, deep learning and visual details seem to have a bright future. Almost every new year brings a new breakthrough in the application of deep learning in our daily lives.