

Deep Learning-Based Knee Osteoarthritis Severity Classification from X-Ray Images Using Transfer Learning and Web-Based Decision Support

Gajanana Narayana Naika

Department of E&CE

KLS VEDIT, Haliyal,

Karnataka, India

Email: naikgajanana21@gmail.com

Nikhil A. Kulkarni

Assistant Professor

Department of E&CE

KLS VEDIT, Haliyal, Karnataka, India

Email: nak@klsvedit.edu.in

Abstract—Knee osteoarthritis affects the knee joint slowly and may reduce comfort, walking ability, and daily movement. In many clinical and academic screening situations, X-ray images are examined to understand how much the joint has changed. The final grade, however, still depends on visual assessment, and small radiographic changes may lead to different opinions. This paper presents a deep learning-based system for classifying the severity of knee osteoarthritis from X-ray images. The system predicts five Kellgren-Lawrence categories: Healthy, Doubtful Osteoarthritis, Minimal Osteoarthritis, Moderate Osteoarthritis, and Severe Osteoarthritis. MobileNetV2 is used as the transfer learning backbone because it is lightweight and suitable for a web-based prototype. The training workflow includes image preprocessing, augmentation, class weighting, focal loss, validation monitoring, checkpoint saving, and early stopping. The application also displays confidence score, class-wise probability values, uncertain-case indication, and a simple rule-based progression note. A Flask web application is developed for user registration, login, X-ray upload, prediction display, and history storage. Testing on 1,656 images produced 55.31% accuracy and a macro F1-score of 52.12%. The model gives better results for visually clear cases such as Healthy and Severe Osteoarthritis, while nearby grades remain difficult because their X-ray patterns are similar. The work presents an academic decision-support prototype for assistive osteoarthritis screening.

Index Terms—Knee osteoarthritis, X-ray classification, deep learning, transfer learning, MobileNetV2, EfficientNetB0, Kellgren-Lawrence grading, medical image classification, Flask, clinical decision support.

I. INTRODUCTION

Knee osteoarthritis is a progressive joint disease that affects cartilage, bone surfaces, and the surrounding structures of the knee. As the condition advances, a person may experience pain, stiffness, and difficulty in movement. Early recognition is useful because timely medical advice, exercise planning, lifestyle changes, and suitable care can help control symptoms and slow further progression.

X-ray imaging is commonly used to examine the severity of knee osteoarthritis. The Kellgren-Lawrence grading system describes the condition from Grade 0 to Grade 4 [1]. Grade 0 represents a normal knee, while Grade 4 represents severe osteoarthritis. Although this grading method is widely used,

it depends on expert visual interpretation. In mild cases, the same X-ray may be read differently by different observers.

Deep learning has become useful in medical image analysis because convolutional neural networks can learn image patterns directly from examples [2], [3]. Medical datasets are often smaller than general image datasets, so training a deep network from the beginning is not always practical. Transfer learning helps in such cases by adapting a model trained on a large image collection to a medical classification task [4]. MobileNetV2 is appropriate for this project because it provides useful feature extraction with lower computational cost [5].

This work develops a knee osteoarthritis severity classification system using transfer learning. The system accepts knee X-ray images and classifies them into five Kellgren-Lawrence grades: Healthy, Doubtful Osteoarthritis, Minimal Osteoarthritis, Moderate Osteoarthritis, and Severe Osteoarthritis. The trained model uses MobileNetV2 along with preprocessing, augmentation, class weighting, focal loss, confidence checking, and Flask-based deployment. SQLite is used to store prediction history, and the result page displays class-wise probabilities for better interpretation.

The main contributions of this work are:

- A five-class deep learning pipeline is developed for knee osteoarthritis severity prediction from X-ray images.
- MobileNetV2 transfer learning is applied for radiographic image classification.
- Class imbalance is handled using class weights and focal loss [7].
- Confidence-aware prediction is included to indicate uncertain outputs.
- A Flask web application is built with authentication, X-ray upload, result display, probability output, and history storage.
- The trained model is evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis.

The rest of the paper is organized as follows. Section II reviews related studies. Section III explains the dataset and preprocessing. Section IV describes the proposed methodology. Section V presents the algorithm. Section VI explains the system architecture. Section VII describes the implementation.

Section VIII reports the experimental results. Section IX discusses the observations. Section X lists the limitations. Section XI presents future work, and Section XII concludes the paper.

II. RELATED WORK

Several studies have used deep learning and transfer learning for medical image classification, including knee osteoarthritis assessment. Previous work shows that CNN-based models can support radiographic analysis. Still, early-stage osteoarthritis and neighbouring grades are difficult to separate because their visual differences are small. Table I summarizes the related studies considered in this work.

The reviewed studies indicate that deep learning can assist knee osteoarthritis classification. However, identifying early grades and separating neighbouring grades remain challenging. Therefore, this work combines transfer learning, imbalance handling, confidence-aware prediction, and web deployment to create a practical assistive system.

III. DATASET AND PREPROCESSING

A. Dataset Description

The dataset used in this work is arranged into training, validation, testing, and auto-test folders. Each split contains five folders labelled from 0 to 4. These labels represent the Kellgren-Lawrence grades used for knee osteoarthritis severity classification [1].

TABLE II
KELLGREN-LAWRENCE GRADE LABELS

Grade	Class Label
0	Healthy
1	Doubtful Osteoarthritis
2	Minimal Osteoarthritis
3	Moderate Osteoarthritis
4	Severe Osteoarthritis

The dataset distribution is shown in Table III. The number of images is not equal across all classes. Grade 4 has fewer samples than Grade 0 and Grade 2. This imbalance may make the model learn majority classes more strongly.

TABLE III
DATASET DISTRIBUTION ACROSS DIFFERENT SPLITS

Split	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4	Total
Train	2286	1046	1516	757	173	5778
Validation	328	153	212	106	27	826
Test	639	296	447	223	51	1656
Auto Test	604	275	403	200	44	1526

B. Image Preprocessing

Each uploaded image is prepared before prediction. The X-ray image is read using OpenCV [13]. Since OpenCV loads the image in BGR order, it is converted into RGB format. The image is then resized to 224×224 pixels to match the MobileNetV2 input size [5].

The preprocessing flow is:

- 1) Load the uploaded knee X-ray image.
- 2) Convert the image from BGR format to RGB format.
- 3) Resize the image to 224×224 pixels.
- 4) Scale or normalize the pixel values according to the model requirement.
- 5) Convert the processed image into a batch tensor.
- 6) Provide the tensor as input to the trained model.

C. Data Augmentation

Data augmentation is used during training to help the model handle image variations. Medical images may vary in brightness, scale, alignment, and acquisition condition. Augmentation creates modified versions of training images, helping the model learn more stable features.

The augmentation operations include:

- Random horizontal flip.
- Random rotation.
- Random zoom.
- Random contrast change.

D. Class Imbalance Handling

The dataset has fewer severe osteoarthritis samples than healthy and minimal osteoarthritis samples. To reduce this imbalance effect, class weights are used during training. Classes with fewer images receive larger weights so that their errors have more influence during learning.

Focal loss is also used along with class weighting. It increases the importance of hard-to-classify examples and reduces the effect of easy examples [7]. This helps when some classes dominate the dataset.

IV. PROPOSED METHODOLOGY

A. Overview of Proposed Method

The proposed method classifies knee X-ray images into five osteoarthritis severity grades. The workflow includes image upload, validation, preprocessing, model prediction, confidence checking, result interpretation, and report storage.

The workflow of the system is:

- 1) The user uploads a knee X-ray through the web page.
- 2) The system checks whether the uploaded file is valid.
- 3) The image is preprocessed and resized.
- 4) The trained MobileNetV2 model generates class probabilities.
- 5) The class with the highest probability is selected.
- 6) Confidence score and probability margin are calculated.
- 7) If the confidence is low, the result is treated as uncertain.
- 8) The grade, confidence score, probabilities, and suggestions are shown to the user.
- 9) The prediction report is saved in the database.

B. Model Architectures

The project supports three model options: a custom CNN, MobileNetV2, and EfficientNetB0. CNN-based methods are commonly used in medical image analysis [3]. The saved model metadata shows that MobileNetV2 is used as the final model.

TABLE I
SUMMARY OF RELATED WORK

Reference	Method / Approach	Major Contribution	Limitation / Research Gap
Kellgren and Lawrence [1]	Radiographic grading system	Proposed the Kellgren-Lawrence scale for assessing osteoarthritis severity from radiographs.	The grading depends on expert judgement and may vary across observers.
Antony et al. [8]	CNN-based knee joint detection and OA severity assessment	Used CNNs for locating knee joints and classifying osteoarthritis severity.	Adjacent OA grades are hard to distinguish because the visual changes are subtle.
Tiulpin et al. [9]	Deep learning-based knee OA diagnosis	Applied deep learning for osteoarthritis diagnosis using plain radiographs.	The result depends on image quality, dataset preparation, and validation on external data.
Chen et al. [10]	Deep neural network with ordinal loss	Used the ordered relation among OA grades to improve severity grading.	Generalization across different clinical datasets remains a concern.
Sandler et al. [5]	MobileNetV2 lightweight CNN architecture	Introduced an efficient CNN architecture useful for transfer learning and lightweight deployment.	The model was designed for general image tasks and requires adaptation for medical images.

TABLE IV
SUPPORTED MODEL ARCHITECTURES

Architecture	Description
Custom CNN	A baseline convolutional model using convolution, batch normalization, activation, pooling, dropout, and dense layers.
MobileNetV2	A transfer learning model using an ImageNet-pretrained lightweight CNN backbone [5].
EfficientNetB0	A transfer learning model based on compound scaling for efficient image classification [6].

C. MobileNetV2 Transfer Learning

MobileNetV2 is used as the main transfer learning model because it is lightweight and computationally efficient. It uses depthwise separable convolutions and inverted residual blocks to reduce the number of parameters while preserving feature extraction capability [5]. This is useful for a web application where prediction time and deployment simplicity are important.

In this system, MobileNetV2 extracts visual features from the knee X-ray image. A classification head is placed above the backbone to assign the image to one of five osteoarthritis grades. The final layer uses softmax activation to produce class probability values.

Transfer learning is suitable for this work because medical datasets are generally smaller than large image datasets [3], [4]. A pretrained model can reuse general visual features and adapt them to knee osteoarthritis classification.

TABLE V
MODEL INPUT CONFIGURATION

Parameter	Value
Input Image Size	224 × 224
Channels	RGB, 3 Channels
Number of Classes	5
Output Activation	Softmax
Optimizer	Adam

D. Focal Loss and Class Imbalance Handling

The dataset is imbalanced, especially for Grade 4 severe osteoarthritis. If the imbalance is ignored, the model may prefer majority classes during training. To reduce this issue, class weights are used. The minority classes receive higher weights so that the model pays more attention to them.

Focal loss is used to improve learning from difficult samples. Although it was introduced for dense object detection, it is also useful for imbalanced classification problems [7]. It reduces the contribution of easy samples and increases the contribution of harder samples.

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (1)$$

Here, p_t is the predicted probability for the actual class, α_t is the class weighting factor, and γ is the focusing parameter. This loss helps the model focus on minority and confusing classes during training.

E. Confidence-Aware Prediction

The system returns more than the predicted label. It calculates the confidence score and probability margin. The confidence score is the highest class probability. The probability margin is the difference between the highest and second-highest probability.

When the confidence score is low or the probability margin is small, the result is marked as uncertain. This is important in medical image applications because not every X-ray image contains clear visual evidence for reliable automatic grading.

F. Rule-Based Progression Outlook

The web system includes a simple rule-based progression outlook. This module provides user-level guidance based on the predicted severity grade. It is not a clinical progression model and does not use longitudinal patient data. It is included only to provide basic educational interpretation with the prediction result.

V. ALGORITHM

The complete prediction procedure is shown in Algorithm 1.

Algorithm 1 Knee Osteoarthritis Severity Prediction

```

1: Start
2: Upload knee X-ray image
3: Validate the image format
4: Read the image using OpenCV
5: Convert the image from BGR to RGB
6: Resize the image to 224 × 224
7: Convert the image into a model-ready tensor
8: Load the trained MobileNetV2 model
9: Predict class probabilities through softmax output
10: Select the class with the maximum probability
11: Calculate the confidence score
12: Calculate the margin between the top two probabilities
13: if confidence is low or margin is small then
14:   Mark the prediction as uncertain
15: else
16:   Assign the predicted Kellgren-Lawrence grade
17: end if
18: Generate class-wise probability output
19: Generate rule-based progression outlook
20: Store the prediction result in the database
21: Display the output to the user
22: Stop

```

VI. SYSTEM ARCHITECTURE

The web-based prediction workflow is shown in Fig. 1. The user uploads an X-ray image through the Flask interface. The image is preprocessed and passed to the MobileNetV2 model. The prediction module returns the grade, confidence value, probabilities, and interpretation to the result page.

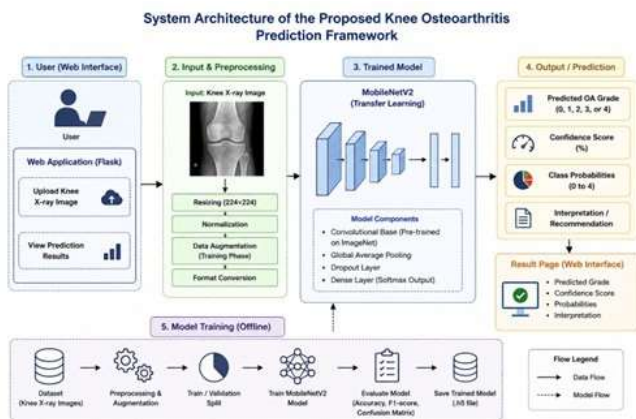


Fig. 1. System architecture of the proposed knee osteoarthritis prediction framework.

VII. SYSTEM IMPLEMENTATION

A. Web Application Design

The proposed system is developed as a Flask-based web application. Flask is selected because it is lightweight and can

be connected easily with Python-based machine learning models [14]. Bootstrap is used to create a simple and responsive interface.

TensorFlow is used for model development and inference [12]. OpenCV is used for reading and preprocessing X-ray images [13]. Flask manages the web routes and deployment workflow [14].

The web application provides:

- User registration and login.
- Knee X-ray image upload.
- File validation for JPG, JPEG, and PNG images.
- Deep learning-based prediction.
- Confidence score and probability display.
- Care suggestion and progression outlook.
- Prediction history storage using SQLite.

B. Application Modules

The main implementation modules are listed in Table VI.

TABLE VI
WEB APPLICATION MODULES

Module	Purpose
app/app.py	Handles routes, authentication, image upload, and prediction history.
prediction/predict.py	Loads the trained model, performs inference, applies confidence logic, and formats the output.
prediction/progression.py	Generates the rule-based progression outlook.
preprocessing/preprocess.py	Converts the uploaded image into the required model input format.
models/train_model.py	Contains the training workflow.
models/evaluate.py	Evaluates the trained model.
utils/metrics.py	Calculates accuracy, precision, recall, F1-score, and confusion matrix.
utils/config.py	Stores paths, class labels, thresholds, and configuration values.

C. Database Storage

SQLite is used to store user information and prediction history. Keeping prediction history allows users to review previous reports and improves the usability of the decision-support prototype.

VIII. EXPERIMENTAL RESULTS

A. Training Configuration

The model is trained using TensorFlow image datasets [12]. The training configuration includes augmentation, class weights, focal loss, validation macro-F1 monitoring, checkpointing, early stopping, and learning-rate reduction on plateau. Fine-tuning of selected final layers is also supported.

The saved model metadata shows MobileNetV2 as the trained architecture. The validation macro-F1 value stored in the metadata is 0.5343.

TABLE VII
SAVED MODEL METADATA

Item	Value
Model Architecture	MobileNetV2
Validation Macro-F1	0.5343
Number of Classes	5
Class Names	0, 1, 2, 3, 4

B. Overall Test Results

The trained model is tested on 1,656 images. Accuracy, precision, recall, F1-score, and confusion matrix are used for evaluation [15]. Table VIII shows the overall results.

The model achieved 55.31% accuracy and a macro F1-score of 52.12%. The weighted F1-score is 54.49%. These values show moderate performance. The model correctly classifies many samples, but it still struggles with neighbouring severity grades.

TABLE VIII
OVERALL TEST EVALUATION RESULTS

Metric	Value
Accuracy	0.5531
Macro Precision	0.5901
Macro Recall	0.4905
Macro F1-score	0.5212
Weighted Precision	0.5658
Weighted Recall	0.5531
Weighted F1-score	0.5449

C. Class-Wise Performance

The class-wise evaluation is shown in Table IX.

TABLE IX
CLASS-WISE PERFORMANCE ON TEST SET

Class	Precision	Recall	F1	Support
Healthy	0.64	0.82	0.72	639
Doubtful OA	0.26	0.30	0.28	296
Minimal OA	0.55	0.41	0.47	447
Moderate OA	0.73	0.41	0.52	223
Severe OA	0.76	0.51	0.61	51

The model performs better for Healthy and Severe Osteoarthritis classes. These two categories are easier to separate visually than the middle grades. The Healthy class achieved an F1-score of 0.72, and the Severe class achieved an F1-score of 0.61. The Doubtful OA class has the lowest F1-score because early-stage features can overlap with healthy or minimal OA images.

D. Confusion Matrix

The confusion matrix for the test set is shown in Table X.

TABLE X
CONFUSION MATRIX ON TEST SET

Actual / Predicted	G0	G1	G2	G3	G4
G0	527	94	18	0	0
G1	172	89	32	3	0
G2	112	140	183	12	0
G3	7	26	91	91	8
G4	0	0	7	18	26

The confusion matrix shows that many Grade 1 images are predicted as Grade 0 or Grade 2. This is expected because doubtful osteoarthritis may not show clear radiographic changes. Grade 2 is also confused with Grade 1, which shows the overlap between adjacent classes. Grade 4 has fewer samples, but the model identifies several severe cases correctly because severe osteoarthritis has more visible structural changes.

IX. DISCUSSION

The experimental results show that the model gives moderate classification performance. The system works better for visually distinct classes, particularly Healthy and Severe Osteoarthritis. Healthy X-ray images generally show normal joint structure, while severe cases show stronger structural changes.

The main difficulty is the separation of adjacent severity grades. The difference between Grade 1 and Grade 2 can be small, and even manual interpretation may differ in such cases. Grade 2 and Grade 3 can also share similar features depending on joint space narrowing, osteophytes, and image quality. This explains the lower recall values observed for some middle grades.

Class imbalance also affects the model. Grade 4 has only 173 training images, which is much lower than the number of Grade 0 and Grade 2 images. Class weighting and focal loss help reduce this issue, but a more balanced dataset would likely improve learning. Independent clinical validation is also needed before considering real-world use.

The Flask web application improves usability by allowing image upload and result viewing through a simple interface. Confidence-aware prediction is useful because it marks uncertain cases instead of forcing a final grade for every image. This is important in medical decision-support applications.

X. LIMITATIONS

The proposed system has some limitations. First, the model performance is moderate and is not sufficient for use as an independent clinical diagnostic tool. Second, early and neighbouring osteoarthritis grades are difficult to classify because their visual differences are small. Third, the dataset is imbalanced, especially for Grade 4. Fourth, the progression outlook is rule-based and does not use longitudinal patient data. Fifth, the system has not been validated by radiologists. Finally, independent clinical datasets are required to evaluate generalization.

XI. FUTURE WORK

The system can be improved by using a larger and more balanced dataset. Stronger architectures such as DenseNet, ResNet, Vision Transformers, or ensemble models can also be tested. Knee joint localization or segmentation may be added before classification so that the model focuses on the relevant joint region. Grad-CAM explainability can be included to show the image regions that influence the prediction [11]. Future versions may also include radiologist-reviewed labels, external clinical validation, and true longitudinal progression prediction.

XII. CONCLUSION

This paper presented a deep learning-based knee osteoarthritis severity classification system using transfer learning and web-based decision support. The system classifies knee X-ray images into five Kellgren-Lawrence grades using a MobileNetV2-based model. The training pipeline includes augmentation, class weighting, focal loss, checkpointing, and validation monitoring. The system also provides confidence-aware prediction, class probability output, and rule-based progression guidance.

A Flask web application was developed to support X-ray upload, prediction result viewing, and prediction history access. Testing on 1,656 images achieved an accuracy of 55.31% and a macro F1-score of 52.12%. The model performs better for Healthy and Severe Osteoarthritis classes, while adjacent grades remain difficult. The system is suitable as an academic prototype and educational assistive tool, but further clinical validation and performance improvement are necessary before practical medical deployment.

REFERENCES

- [1] J. H. Kellgren and J. S. Lawrence, "Radiological assessment of osteoarthritis," *Annals of the Rheumatic Diseases*, vol. 16, no. 4, pp. 494–502, 1957.
- [2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
- [3] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 6105–6114.
- [7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [8] J. Antony, K. McGuinness, K. Moran, and N. E. O'Connor, "Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks," in *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2017, pp. 376–390.
- [9] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Scientific Reports*, vol. 8, no. 1, pp. 1–10, 2018.
- [10] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 84–92, 2019.
- [11] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [12] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
- [13] G. Bradski, "The OpenCV library," *Dr. Dobbs's Journal of Software Tools*, 2000.
- [14] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed. O'Reilly Media, 2018.
- [15] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009.