Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

YOLOv8-Assisted Knee Osteoarthritis Severity Assessment from X-ray Images

Dhanya K.R.

Assistant Professor, Department of Computer Science with Artificial Intelligence and Data Science, KPR College of Arts, Science and Research, Coimbatore, India.

E-mail: dhanya.kr@kprcas.ac.in

Abstract

Knee osteoarthritis is a progressive degenerative joint disorder characterized by cartilage deterioration, leading to pain, stiffness, reduced mobility, and functional impairment. Early detection and accurate grading of disease severity play a crucial role in effective clinical decision-making and prevention of further joint damage. Most existing deep learning approaches focus mainly on image-level classification using conventional CNN architectures and do not provide precise localization of the affected knee region, which is essential for clinical interpretation. In addition, limited research has explored unified object detection frameworks for simultaneous localization and severity grading of knee osteoarthritis. To address these limitations, this study proposes a YOLOv8-based deep learning framework for automated knee osteoarthritis detection and severity classification using radiographic X-ray images collected from a publicly available Kaggle dataset. The dataset is carefully annotated using Label Studio by marking the knee joint region and assigning five severity grades: Healthy, Doubtful, Minimal, Moderate, and Severe. The proposed model is trained using GPU acceleration to enhance computational efficiency and improve detection performance. Experimental evaluation is performed using standard metrics and the proposed system achieves a mAP@0.5 of 85.2%, demonstrating strong capability in both accurate localization and classification of knee osteoarthritis severity. It can be deployed in clinical decision support systems, radiology departments, and telemedicine platforms to enable rapid knee X-ray screening, providing radiologists, orthopaedic specialists, and physicians with an automated second opinion that reduces diagnostic time and human variability.

DOI: https://doi.org/10.36548/jaicn.2026.3.004
Received: 18.05.2026, received in revised form: 22.06.2026, accepted: 10.07.2026, published: 21.07.2026
© 2026 Inventive Research Organization. This is an open access article under the Creative Commons Attribution-NonCommercial International (CC BY-NC 4.0) License

Keywords: Knee Osteoarthritis, YOLOv8, Deep Learning, X-ray Image Analysis, Medical Image Classification.

1. Introduction

Knee osteoarthritis (OA) ranks the second highest in terms of musculoskeletal problems that plague many millions of individuals across the world [1]. Knee osteoarthritis is a degenerative disease of the joints whereby there is gradual wear and tear on the joints, inflammation of the affected joint, and alteration of the structure of the bone near the joint [2]. The disease can get worse to a point where the patient feels pain, stiffness, and even difficulty in mobility. This type of knee osteoarthritis is common among old people, although it may also affect young people due to obesity, previous injury, genetic issues, and excessive strain on joints. It affects women more than men and occurs more often in older individuals [3]. X-rays serve one of the ways to diagnose the severity of knee osteoarthritis. In order to make further diagnosis and treatment easier, severity of knee OA can be assessed using 5 types which include Healthy (0), Doubtful (1), Mild (2), Moderate (3), and Severe (4). These categories help medical practitioners find the disease stage sort of faster and then decide on treatment directions, like lifestyle changes, physical therapy, or surgical operations such as joint replacement.

Doctors used to diagnose and rate knee OA in the past manually from the patients' X-rays images. This is a very time-consuming task, as there would be many differing diagnoses due to various interpretations of the level of severity made by various specialists [4]. Consequently, it becomes very difficult to give a correct assessment of the disease. As a result, treatment is delayed, inaccurate, or even incorrect. There is a necessity for an automatic and accurate device which gives out consistent diagnoses. There has been a radical shift in the medical imaging through the development of computerized algorithms called deep learning models that can analyze the information provided by the imaging technology with high accuracy and speed at a level close to that of expert medical practitioners. Most of the literature does not focus on identifying the regions of the joint where the progression of the disease is highest but rather only categorizes the degree of progression of the disease [5]. Thus, with an aim to address the challenges, this study recommends an enhanced deep learning-based method for the diagnosis and grading of knee osteoarthritis. The main contributions made by this study are outlined below:

- Adopts YOLOv8 model as the foundation for the knee osteoarthritis detection and severity assessment approach based on radiological images obtained from Kaggle datasets, enhancing both detection and classification processes.
- Uses Label Studio tool for the preparation of the datasets through labeling knee X-rays with bounding boxes to be categorized into five different levels: Healthy, Doubtful, Minimal, Moderate, and Severe.
- Uses YOLOv8 as a deep learning model to detect and classify the severity levels of the knee osteoarthritis condition with GPU capability.
- Evaluates the developed YOLOv8 model using various metrics including Precision, Recall, F1 Score, and mAP@0.5.

2. Related Works

Table 1. Comparative summary of existing approaches for knee OA detection

Author	Method	Accuracy	Limitations
Chen et al. [6]	YOLOv2 + VGG-19 + Ordinal Loss	69.70%	Limited to KL grading only, poor performance on early-stage grades (KL-1, KL-2)
Tiulpin et al. [7]	Deep Siamese CNN + ResNet-34	66.71%	Two-branch architecture increases computational cost, poor generalization
Xiaolu Ren [8]	VGG-based joint center extraction + ResNet-50 classification	81.41%	Struggles to accurately distinguish early OA grades (KL-1, KL-2, and KL-3)
Hae Jeong Lee [9]	Ensemble CNN	76.93%	High ensemble computational cost, no localization module, and weak early-grade detection
Saman Shahid [10]	DenseNet-121	68.85%	Dataset imbalance and poor KL 0–1 differentiation
Cueva [11]	ResNet-34	61%	KL-1/KL-2 misclassification
Zhong [12]	YOLOv8s	79.69 %	Hard to catch those subtle early-stage KOA signs because the x-ray patterns overlap, even when the disease is still in a mild grade.

Based on the analysis of the existing techniques as shown in Table 1, it can be seen that there are various techniques that lack efficiency in terms of identifying knee osteoarthritis, particularly those with five classifications according to the KL grading method. Moreover, most techniques find it difficult to detect the earlier stages of osteoarthritis (KL-1 and KL-2), given the similarities between the classes. The other challenge is that there are some techniques that uses multi-stage frameworks, resulting in increased computational costs. Another challenge that is common in many systems is dataset imbalance. To overcome these challenges, the new system employs YOLOv8, a single-stage technique that involves localization and classification simultaneously.

3. Proposed Work

The proposed methodology of the knee osteoarthritis detection system is illustrated in Figure 1. It begins with the collection of knee X-ray images, followed by selection of high-quality samples for effective training.

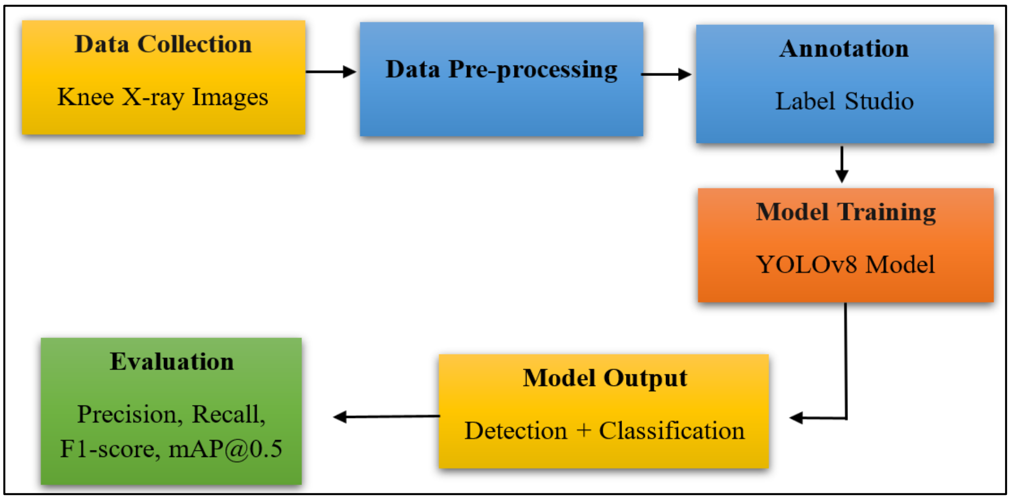

Figure 1. Workflow of the proposed system

Images are annotated using Label Studio by labeling the knee joint region and severity categories. This dataset is then used for training YOLOv8 to perform detection and classification at the same time. The trained model is then used for prediction of osteoarthritis severity, along with the location of the affected knee region.

3.1 Data Collection

For this study, a total of 3,137 knee X-ray images were collected from a Kaggle knee osteoarthritis dataset [6]. The dataset contains radiographic images categorized into five osteoarthritis severity classes based on standard clinical grading. These classes representing different stages of cartilage degeneration and joint damage as shown in Table 2.

Table 2. Total number of images per class

Class (Severity)	Total Images
0 (Healthy)	650
1 (Doubtful)	802
2 (Minimal)	740
3 (Moderate)	650
4 (Severe)	295

3.2 Data Pre-processing

Prior to training, all images for the knee joint are pre-processed to ensure that they are standardized and the knee joint structures are easily detected through the YOLOv8 algorithm. This process begins with the resizing of all images to a size of 640 × 640 pixels, as required by YOLOv8, as indicated in Eq. (1).

$$I_i^{resized} = \text{Resize}(I_i, 640{,}640) \tag{1}$$

After scaling, the pixel values are normalized from 0-255 to 0-1 to reduce variance in illumination and make the training process more stable. This process of normalization can be seen in Eq. (2).

$$\hat{I}_i(x, y) = \frac{I_i^{resized}(x,y)}{255} \tag{2}$$

Then, Contrast Limited Adaptive Histogram Equalization (CLAHE) is performed to enhance the contrast of knee X-ray images locally and bring into focus the important details that have been identified. This enhancement process is defined in Eq. (3).

$$I_i^{CLAHE}(x, y) = \text{CLAHE}_{T,c}\left(\widehat{I_i}(x, y)\right) \tag{3}$$

ISSN: 2582-2012

The pre-processed images are then utilized in annotation processes and training of the proposed knee osteoarthritis detection and classification model based on the YOLOv8 architecture.

In order to solve the problem of natural class imbalance in the training data, a weighting of loss is used in the training process. The weights of individual classes depend on the inverse proportionality to their numbers, ensuring that classes with fewer samples are sufficiently considered in the learning phase. Class weights are specified by Eq. (4).

$$W_c = \frac{N_{total}}{N_c} \tag{4}$$

where N_{total} is the total number of training images, and N_c refers to the number of images in class c. The method increases the weight of penalties for incorrectly classifying the smaller classes, resulting in a better detection rate of all severity levels.

3.3 Data Annotation

Here, for each of the knee X-ray images, we use I_i, to represent the image, where $i = 1,2,\ldots,N$. For each image, the region of the knee joint is annotated with a bounding box. The bounding box can be formulated as Eq. (5):

$$B_i = (x_i, y_i, w_i, h_i) \tag{5}$$

where x_i and y_i are the coordinates for the top-left corner of the bounding box, and $w_i$and h_i represent its dimensions. After obtaining the bounding box for an image, it is given an osteoarthritis grade label C_i based on the equation as defined in Eq. (6):

$$C_i \in \{0,1,2,3,4\} \tag{6}$$

with 0 denoting Healthy, 1 denoting Doubtful, 2 denoting Minimal, 3 denoting Moderate, and 4 denoting Severe. Finally, the entire set of annotated data can be described as an image-bounding box-label tuple, as shown in Eq. (7):

$$D = \{(I_i, B_i, C_i)\}_{i=1}^{N} \tag{7}$$

So, this dataset formulation is kind of used, for training the proposed YOLOv8 model for detecting knee osteoarthritis and also for its classification, in general.

3.4 Model Selection and Training

The YOLOv8 model architecture will be used in the design of the proposed system. The architecture is a state-of-the-art object detector that is known for its speed and accuracy in detecting objects. The YOLOv8 object detector comprises three main parts; the Backbone, Neck, and Detection head. In the YOLOv8 architecture, the backbone plays an essential role of extracting features from the input knee X-ray images using down-sampling and convolutions techniques. This part learns low and high-level features of bone edges, narrowed joints, osteophyte formations, and deformities linked to osteoarthritis severity. The extracted feature maps are further forwarded to the Neck Module, which does multi-scale feature fusion. This is necessary since osteoarthritis may manifest in both small-scale and large-scale features, where small-scale features include minor bone spurs while large-scale features include joint misalignment or even significant narrowing. The architectural design of the suggested YOLOv8-based knee osteoarthritis detection system is shown in Figure 2 below, illustrating the feature processing from Backbone Network via CSPDarknet53 for feature extraction, followed by the Detection Head with an anchor-free prediction approach for bounding box prediction and OA classification.

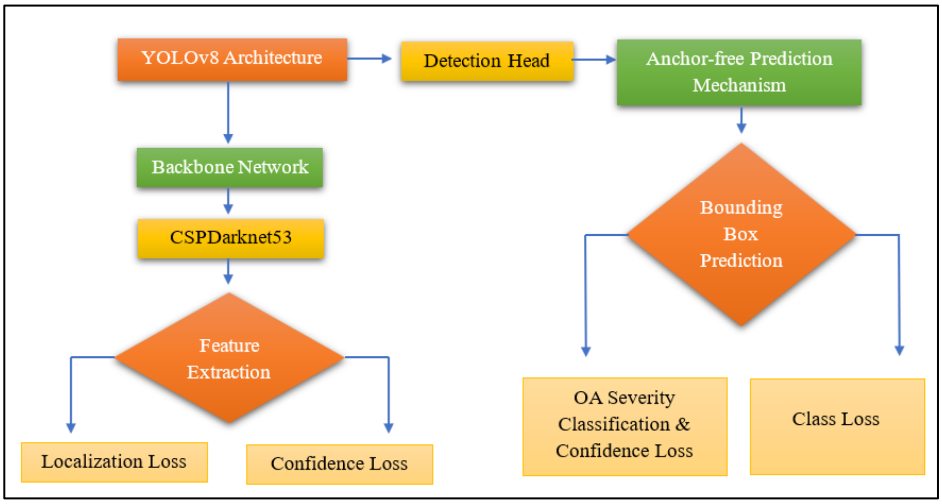

Figure 2. YOLOv8 architecture for knee osteoarthritis detection and severity classification

As a result of using characteristics from multiple layers, YOLOv8 can better detect the knee joint area. Unlike the older models of YOLOv8, which rely on the use of anchors for predicting the objects, this model relies on the anchor-free object detection approach, thereby facilitating accurate localization regardless of size and orientation differences in the knee joint on radiographic images. Lastly, through the detection head, YOLOv8 predicts the bounding box x, y, w, h, confidence score, and class probability.

The training dataset is split into 80% of the data used as training data and 20% as validation data, which is done in such a way as to include all classes of severity in each split. The images are arranged in different class-based folders, and a YAML file is generated in order to define dataset locations and class names for the training of the YOLOv8 model. Training of the YOLOv8 model is done for 120 epochs, with a batch size of 8. In the training process, the model is optimized with the help of SGD optimizer, with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. Also, a cosine learning rate schedule is used, with 3 epochs of warmup in order to attain stable convergence during training. During training, the model is executed in a GPU-supported environment for speed and performance benefits.

3.5 Model Evaluation

Evaluation of the trained YOLOv8 model is done using various metrics for object detection. The metrics will be used to determine the effectiveness with which the model detects the knee joint area and its classification for osteoarthritis severity. The precision metric will show the percentage of the true positive predictions compared to all the predicted ones while the recall metric shows the capability of the model to detect all the relevant knee joint areas. The evaluation metrics are described below in Eq. (8) and Eq. (9).

$$P = \frac{TP}{TP+FP} \tag{8}$$

$$R = \frac{TP}{TP+FN} \tag{9}$$

In Eq. (8), TP represents true positives and FP represents false positives, while in Eq. (9), FN represents false negatives. The Average Precision (AP) for each class is computed as the area under the Precision–Recall curve, as defined in Eq. (10).

$$AP_i = \int_0^1 P(R)\, dR \tag{10}$$

In Eq. (10), $P(R)$ represents the Precision–Recall curve for the i^{th} class, where $i = 1,2,\dots,N$. The Mean Average Precision (mAP) across all classes is defined as the average of AP values over all severity classes, as shown in Eq. (11).

$$mAP = \frac{1}{N}\sum_{i=1}^{N} A\,P_i \tag{11}$$

In Eq. (11), N represents the total number of osteoarthritis severity classes, and AP_i represents the Average Precision of the i^{th} class. The final value of mAP shows the performance of the proposed model YOLOv8 on the object detection and classification task. The test is performed using the validation dataset (20% of total data), and it includes a separate evaluation for each class (from Healthy to Severe).

4. Results and Discussion

4.1 System Configuration

The suggested architecture of the model was implemented and tested under a GPU-powered computer setup. Model training and validation processes were conducted through Python 3.10 programming language using the Ultralytics YOLOv8 model (v8.2). Image preprocessing and visualization were realized using the OpenCV 4.8 and NumPy 1.24 libraries. The annotations of the X-rays of knees dataset were manually made using Label Studio software (v1.10). The tests were conducted on Windows 10 (64-bit) operating system, featuring an NVIDIA GPU, 16 GB RAM, and adequate storage.

4.2 Performance Analysis

Evaluation was conducted on 627 validation images to determine the detection and classification accuracy of the trained YOLOv8 model. Several graphs related to the evaluation results and prediction were plotted to understand the learning characteristics of the model. The performance of the class-wise evaluation is tabulated in Table 3. The trained model performed the best in terms of detecting and classifying the Healthy class with an mAP@0.5 value of 0.950. As such, the model had a high ability to detect healthy knees. On the other hand, the Doubtful class exhibited poor performance compared to others, which could have resulted from the closeness of severity levels.

Table 3. Class-wise performance metrics

Class	Precision	Recall	mAP@0.5	mAP@0.5 –0.95
Healthy	0.851	0.877	0.950	0.663
Doubtful	0.665	0.785	0.719	0.463

Minimal	0.709	0.837	0.868	0.583
Moderate	0.793	0.798	0.846	0.447
Severe	0.795	0.886	0.876	0.426
Overall	0.763	0.837	0.852	0.516

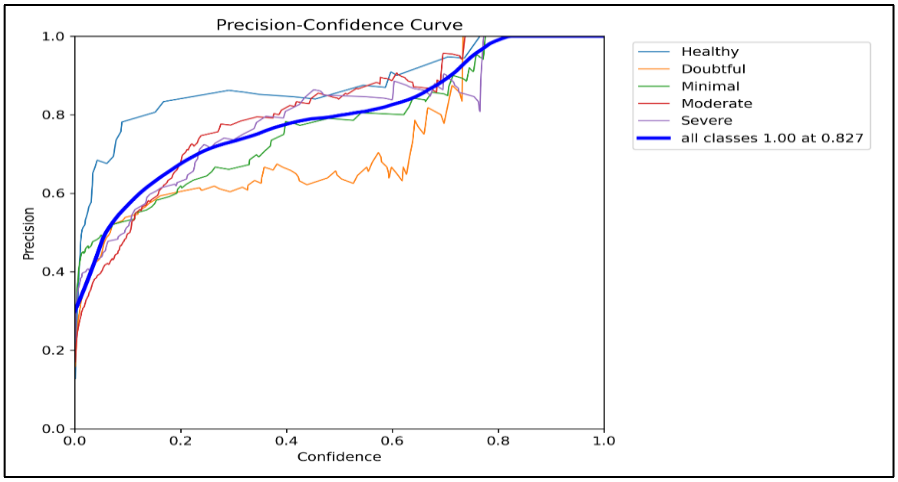

Figure 3. Precision-confidence curve

The Precision-Confidence plot presented in Figure 3 shows the results for all five classes of osteoarthritis severity identified using the YOLOv8 algorithm. It is observed that Healthy class has the highest precision rate amongst all classes in the model irrespective of confidence level. The class of Doubtful has the worst precision rate along with high variability in the precision score due to similarity in appearance with nearby grades of KL. Both Moderate and Severe classes have increasing precision rates with higher values of confidence, thus ensuring good performance at higher confidence levels. All five classes tend towards a precision score of 1.00 when the confidence level is 0.827.

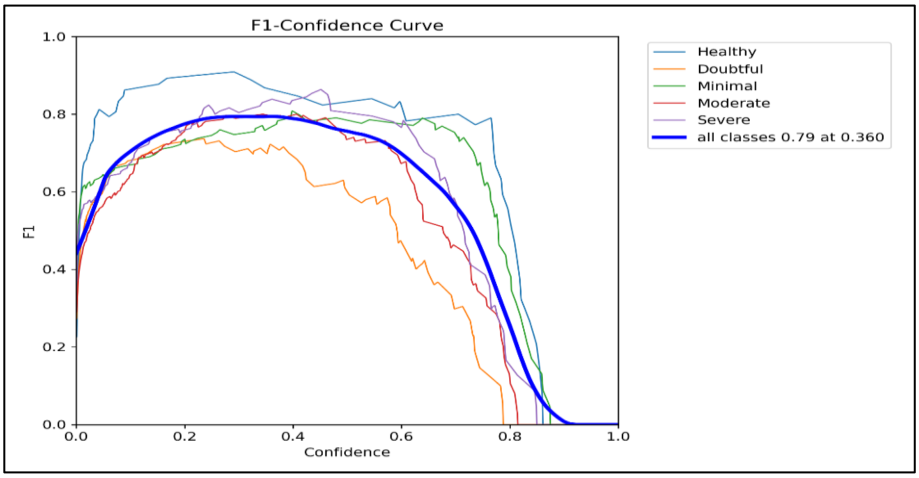

Figure 4. F1-confidence curve

Figure 4 demonstrates the F1-Confidence curve of all five severity classes of osteoarthritis, with the Healthy category having the maximum F1 value because it is the most balanced in terms of precision and recall across all levels of confidence. The Doubtful category retains low F1 values and gradually reduces in the course of analysis, thus proving that it is hard to classify because of the similarity of categories. The Minimal, Moderate, and Severe categories display stable values of F1 in the middle of the confidence level before rapidly falling down at a high confidence level. The overall graph of the F1 score across all categories reaches the maximum value of 0.79 at the confidence level of 0.360.

The normalized confusion matrix (Figure 5) shows the Healthy class having a perfect prediction of 0.96, while the Severe class has a classification accuracy of 0.91 and the Doubtful class an accuracy of 0.87. The Minimal and Moderate classes have accuracies of 0.84 and 0.82, respectively.

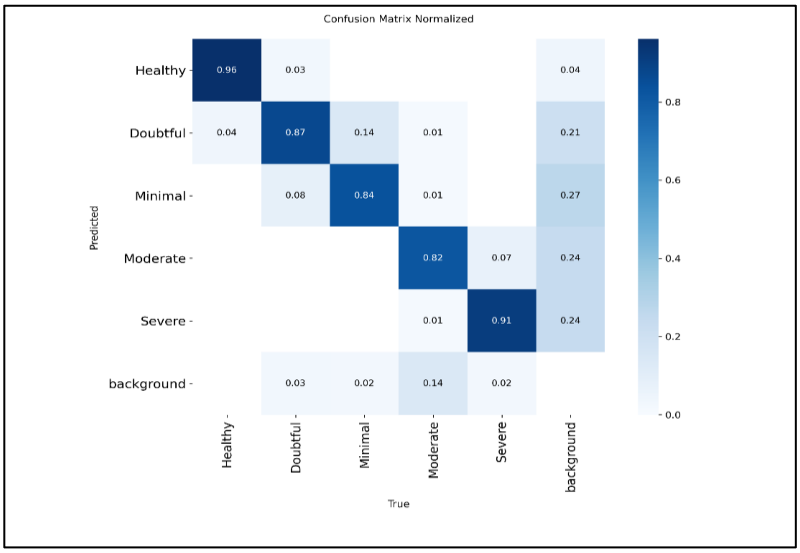

Figure 5. Normalized confusion matrix

The trend in misclassification has been noticed in the column Background, where classes Doubtful, Minimal, Moderate, and Severe have confusion rates of 0.21, 0.27, 0.24, and 0.24, respectively, suggesting that some areas of the knees were not detected correctly. Misclassifications between classes are common from a clinical perspective, given that adjacent classes have radiographic similarities; however, the high rate of diagonal dominance indicates that the suggested YOLOv8 model performs well in classifying OA classes.

The following Figure 6 are the examples of predictions of the trained YOLOv8 model, where bounding boxes have been used to indicate the detected regions of the knee joints, with their respective severity label predictions. The YOLOv8 model is able to localize the region of the knee joint regardless of their orientation or image quality, thus indicating the effectiveness of the model in terms of robust detection of the ROI. The localization is then accompanied by a severity level labels, which implies the success of the model in detecting and classifying the input image.

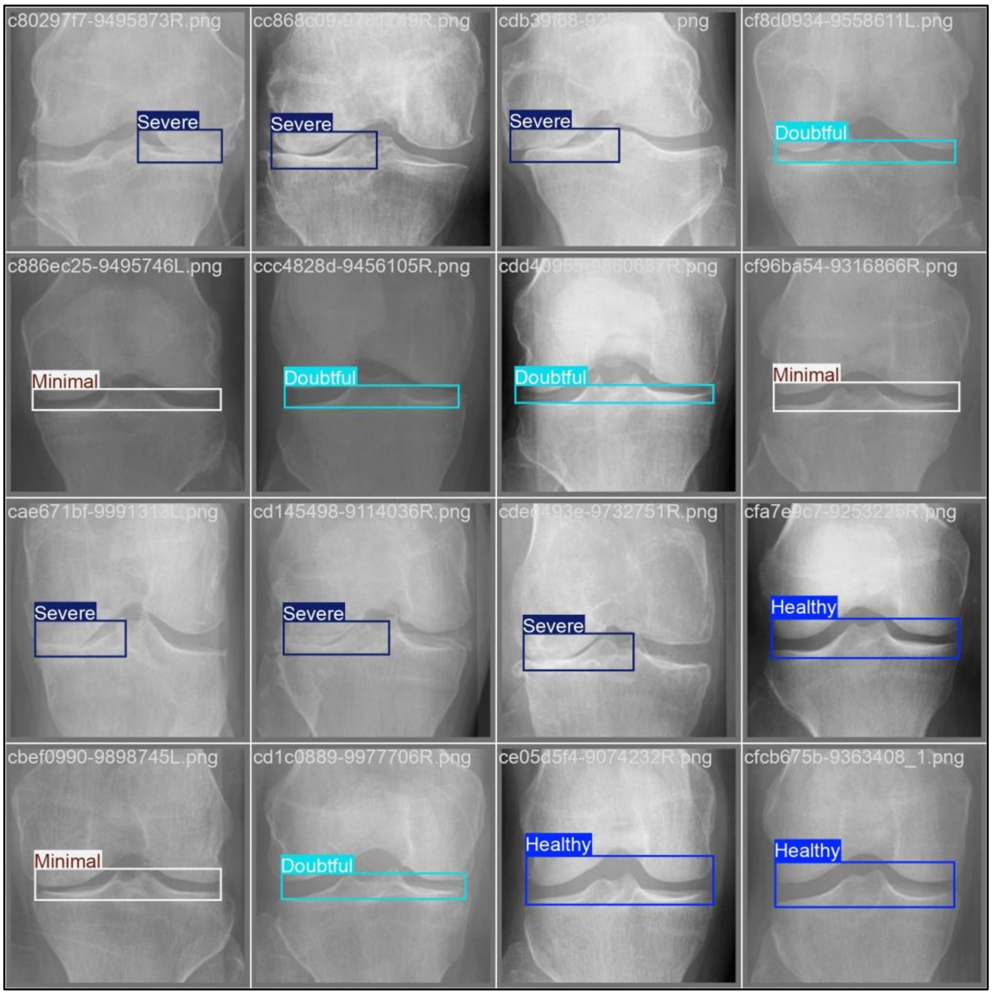

Figure 6. Sample prediction images

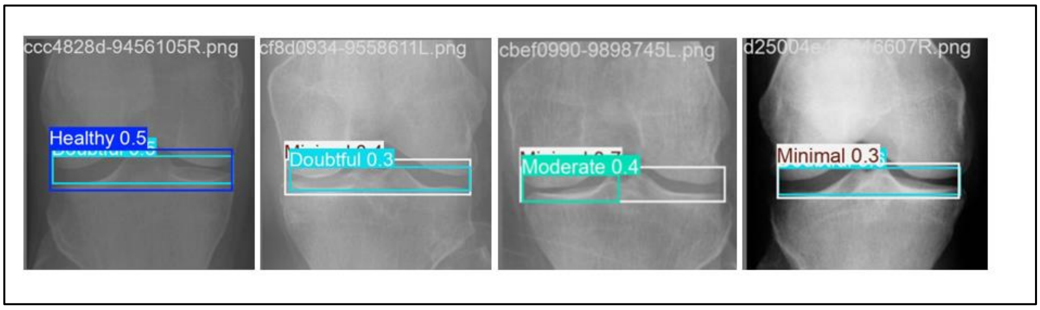

Figure 7. Sample error prediction outputs of the YOLOv8 model

However, there have been some false predictions made by the model are illustrated in Figure 7. False predictions can be occurred in the cases involving neighboring grades like Healthy, Doubtful, Minimal and Moderate, which have low confidence values ranging from 0.3 to 0.5. This is due to the fact that all these grades are visually indistinguishable in x-rays, considering the differences in narrowing of joint spaces and bone changes are extremely minute.

5. Conclusion

In this study, a computerized method for knee joint osteoarthritis detection and grading of its severity has been propossed, which utilizes the state-of-the-art YOLOv8 deep neural network algorithm. The dataset was annotated using Label Studio and then subjected to several preprocessing steps such as resizing, normalizing, and CLAHE. The model was trained for 120 epochs using GPU power and validated. As shown by experimental data, the proposed architecture managed to produce an mAP@0.5 equal to 85.2%, out of which the Healthy class had the best score of 0.950. The additional Precision-Confidence and F1-Confidence curves provide additional validation for the robustness of the model at optimized confidence levels, while the confusion matrix demonstrates correct classification among all five grade categories. Therefore, the developed YOLOv8 approach is proved to be a highly accurate and robust method for automated analysis and classification of knee OA patients. Future research directions for improved performance include integrating transformer models like Vision Transformers for better distinction among grades that are visually similar, multimodal approaches by utilizing multiple modalities (X-ray, MRI, CT), federated learning for cross-hospital training datasets.

References

[1] Steinmetz, Jaimie D., Garland T. Culbreth, Lydia M. Haile, Quinn Rafferty, Justin Lo, Kai Glenn Fukutaki, Jessica A. Cruz et al. "Global, Regional, and National Burden of Osteoarthritis, 1990–2020 and Projections to 2050: A Systematic Analysis for the Global Burden of Disease Study 2021." The Lancet Rheumatology 2023, vol. 5, no. 9: e508-e522.

[2] Thomas, Kevin A., Łukasz Kidziński, Eni Halilaj, Scott L. Fleming, Guhan R. Venkataraman, Edwin HG Oei, Garry E. Gold, and Scott L. Delp. "Automated

Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks." Radiology: Artificial Intelligence 2020, vol. 2, no. 2: e190065.

[3] Jin, Ziyi, Dandan Wang, Huayong Zhang, Jun Liang, Xuebing Feng, Jinkou Zhao, and Lingyun Sun. "Incidence Trend of Five Common Musculoskeletal Disorders from 1990 to 2017 at the Global, Regional and National Level: Results from the Global Burden of Disease Study 2017." Annals of the rheumatic diseases 2020, vol. 79, no. 8: 1014-1022.

[4] Tiulpin, Aleksei, and Simo Saarakkala. "Automatic grading of Individual Knee Osteoarthritis Features in Plain Radiographs Using Deep Convolutional Neural Networks." Diagnostics 2020, vol. 10, no. 11, 932.

[5] Yunus, Usman, Javeria Amin, Muhammad Sharif, Mussarat Yasmin, Seifedine Kadry, and Sujatha Krishnamoorthy. "Recognition of knee Osteoarthritis (KOA) Using YOLOv2 and classification Based on Convolutional Neural Network." Life 2022, vol. 12, no. 8: 1126.

[6] Chen, Pingjun, Linlin Gao, Xiaoshuang Shi, Kyle Allen, and Lin Yang. "Fully Automatic Knee Osteoarthritis Severity Grading Using Deep Neural Networks with a Novel Ordinal Loss." Computerized Medical Imaging and Graphics 2019, vol. 75, 84-92.

[7] Tiulpin, Aleksei, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. "Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach." Scientific reports 2018, vol. 8, no. 1, 1727.

[8] Ren, Xiaolu, Lingxuan Hou, Shan Liu, Peng Wu, Siming Liang, Haitian Fu, Chengquan Li, Ting Li, and Yongjing Cheng. "OA-MEN: A Fusion Deep Learning Approach for Enhanced Accuracy in Knee Osteoarthritis Detection and Classification Using X-Ray Imaging." Frontiers in Bioengineering and Biotechnology 2025, vol. 12, 1437188.

[9] Pi, Sun-Woo, Byoung-Dai Lee, Mu Sook Lee, and Hae Jeong Lee. "Ensemble Deep-Learning Networks for Automated Osteoarthritis Grading in Knee X-ray Images." Scientific Reports 2023, vol. 13, no. 1, 22887.

[10] Shahid, Saman, Aamir Wali, Aatir Javaid, Shahid Zikria, Onur Osman, and Jawad Rasheed. "Potential of AI-based Diagnostic Grading System for Knee Osteoarthritis." Frontiers in Medicine 2025, vol. 12, 1707588.

[11] Cueva, Joseph Humberto, Darwin Castillo, Héctor Espinós-Morató, David Durán, Patricia Díaz, and Vasudevan Lakshminarayanan. "Detection and Classification of Knee Osteoarthritis." Diagnostics 2022, vol. 12, no. 10, 2362.

[12] Zhong, Guoyun, Peng Ding, Jianfeng He, Meifeng Liu, Haibo Sun, Fan Diao, Xueyuan Wang, Lin Ma, and Mengmeng Wang. "KOA-YOLOv8s: An Automatic Diagnostic Algorithm for X-ray Images of Knee Osteoarthritis Based on the Improved YOLOv8s." Journal of King Saud University Computer and Information Sciences 2026, vol. 38, no. 4, 140.