Utilizing Cascade Deep Metric Learning for the Kellgren-Lawrence Grading of Knee Osteoarthritis Classification from X-Ray Images

Supatman
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
7022212011@student.its.ac.id

Eko Mulyanto Yuniarno
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia | Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
ekomulyanto@ee.its.ac.id

Mauridhi Hery Purnomo
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia | Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
hery@ee.its.ac.id (corresponding author)

Received: 12 January 2026 | Revised: 10 February 2026 | Accepted: 20 February 2026

 DOI: https://doi.org/10.48084/etasr.17478

ABSTRACT

Accurate automated grading of knee osteoarthritis from X-ray images remains challenging due to inter-radiologist label noise and the inherent ordinal nature of the Kellgren-Lawrence (KL) grading scale. Most existing deep learning approaches address label noise handling, feature embedding, and ordinal modeling as separate components, which limits robustness and consistency across adjacent grades. This study presents a Cascade Deep Metric Learning (CDML) framework that integrates adaptive label reliability updating and ordinal-aware metric learning within a unified cascade design. By iteratively refining feature embeddings under ordinal constraints, the proposed framework explicitly addresses both noisy annotations and ordinal dependencies in KL grading. Experiments conducted on the Osteoarthritis Initiative (OAI) dataset demonstrated that the proposed method achieved an accuracy of 82.12%, a Mean Absolute Error (MAE) of 0.179, and a Quadratic Weighted Kappa (QWK) of 0.935, outperforming baseline deep metric learning methods. Ablation studies further confirmed the effectiveness of the cascade design in improving robustness and preserving clinically consistent ordinal predictions.

Keywords-cascade deep metric learning; Kellgren–Lawrence grading; knee osteoarthritis; label noise handling; ordinal classification

I. INTRODUCTION

Knee Osteoarthritis (KOA) is a prevalent degenerative joint disease and a significant cause of disability worldwide [1]. Deep learning has been increasingly applied to musculoskeletal disorders, including hip osteoarthritis [2], enabling automated image-based assessments. Radiographic indicators of the progression of KOA include osteophyte formation, joint space narrowing, and subchondral bone changes associated with cartilage degeneration [3]. Despite the exploration of complementary imaging and signal-based approaches [4], conventional radiography remains the main modality for the evaluation of KOA due to its accessibility and cost-effectiveness.

The KL system is the gold standard for the evaluation of KOA, with a scale of 0 to 4 [5]. KL grading relies heavily on expert visual interpretation and is therefore inherently subjective [6]. Substantial inter-observer variability has been reported [7]. Furthermore, intra-observer inconsistency occurs, leading to noisy and inconsistent labeling in practical settings [8]. This ambiguity arises primarily at adjacent levels, such as KL-0, KL-1, and KL-2, due to subtle visual differences [9]. Therefore, publicly available datasets exhibit significant label

ambiguity, posing a major challenge to the development of automated image-based methods.

Deep learning, particularly CNN-based models, has achieved strong performance across diverse medical image classification tasks, including pneumonia detection from chest X-rays [10], hematological disease diagnosis using erythrocyte images [11], and neurological disorder classification using ensemble learning [12]. In the context of osteoarthritis, DL-based approaches have been widely adopted for automated KL grading to enhance objectivity and reproducibility [13]. Nevertheless, most existing methods assume uniform label reliability and model KL grading as a standard multi-class problem, neglecting inter-radiologist noise, visual ambiguity, and the ordinal nature of KL grades [14]. Moreover, label refinement, feature embedding, and ordinal modeling are often handled in isolation, limiting robustness and consistency across adjacent grades.

To address the limitations of existing approaches, this study proposes a CDML framework for automated KL grading that unifies label cleaning, metric embedding, and ordinal modeling in a single iterative pipeline. By enabling mutual reinforcement between embedding refinement and ordinal consistency, the proposed framework improves robustness to label noise and enhances class separability. To our knowledge, this is the first approach to integrate these components into a unified cascade tailored to the ordinal and ambiguity-prone nature of KOA grading. The main contributions of this study are summarized as follows:

- The proposed CDML framework integrates label refinement and ordinal-aware embedding in a unified iterative pipeline for KL grading.
- A cascade mechanism jointly suppresses label noise and enforces ordinal consistency, reducing misclassification across non-adjacent KL grades.
- Experimental results demonstrate consistent improvements in accuracy, MAE, and QWK over baseline deep metric learning methods.

II. STATE-OF-THE-ART IN KNEE OSTEOARTHRITIS CLASSIFICATION

Deep learning-based CNN models have driven significant progress in the automated classification of KL grades in KOA. Early studies using pre-trained CNN architectures demonstrated the feasibility of direct KL grading from knee X-ray images [15] but struggled with the ambiguity and ordinal nature of KL labels. Ordinal-aware learning strategies have been introduced to address this limitation. In [16], an ordinal loss function was employed, achieving 69.70% accuracy and revealing the difficulty of separating adjacent grades. Subsequent CNN-based approaches incorporated bilateral knee information [17], hierarchical classification schemes [18], and ensemble learning [19], reporting accuracies ranging from 65.98 to 76.93% that indicate moderate improvements through architectural refinement.

Recent studies have shifted their focus to label reliability. Confidence-aware sampling preserved ordinal consistency with 70.13% accuracy [20], while data-driven relabeling using EfficientNetB5 and CleanLab achieved 82.07% accuracy [21], highlighting the critical role of label quality in KL grading performance. Nevertheless, prior studies predominantly tackle ordinal modeling, label cleaning, or confidence-aware learning as isolated components. An integrated cascade framework that jointly performs adaptive label cleaning, relabeling, and ordinal-aware metric learning remains largely unexplored. This research gap motivates the proposed CDML framework, which unifies these components to enhance inter-grade discrimination and improve overall KL grade classification accuracy.

III. MATERIALS AND METHODS

This study presents a CDML framework for automated KL grade classification from knee X-ray images as shown in Figure 1. Cascade-I applies VGG19-based [22] embedding with 5-fold cross-validation and Out-Of-Fold (OOF) metric learning for label cleaning and adaptive relabeling, while Cascade-II performs ordinal-aware metric learning-based classification to model KL grade progression. Performance was evaluated using accuracy, MAE, quadratic QWK, and confusion matrices, with Grad-CAM providing model interpretability [23]. All experiments were implemented in Python on a system equipped with an Intel Xeon W-3425 CPU, 32 GB RAM, and NVIDIA RTX A4000 GPU.

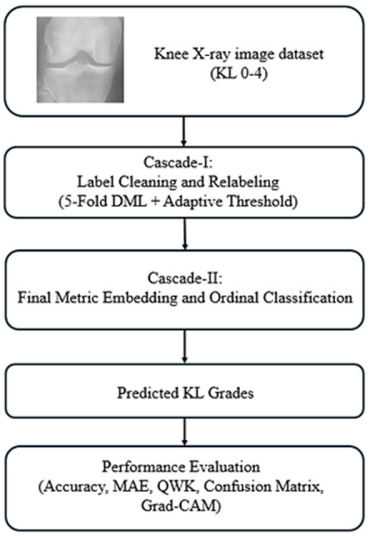

Fig. 1. Block diagram illustrating the overall workflow of the proposed CDML framework.

A. Dataset

The knee X-ray dataset used in this study is publicly available from the OAI through Mendeley Data [24]. The images were annotated according to the KL grading system (grades 0–4), and Table I summarizes the dataset distribution. Further details on data acquisition are provided in the original OAI publications [25]. All data usage complied with the dataset terms. In addition to the internal training, validation, and test splits, the dataset provides an official Auto Test partition, which is reported in Table I for completeness. This study used only the training, validation, and test splits for model development and performance evaluation.

TABLE I. DATASET DISTRIBUTION PER KL GRADE

Split	KL-0	KL-1	KL-2	KL-3	KL-4
Train	2,286	1,046	1,516	757	173
Validation	328	153	212	106	27
Test	639	296	447	223	51
Auto Test	604	275	403	200	44

As summarized in Table I, the dataset comprises 9,786 grayscale X-ray images with a resolution of 224×224 pixels. For model development, images from the training, validation, and test splits were merged by KL class, yielding 3,253 images for KL-0, 1,495 for KL-1, 2,175 for KL-2, 1,086 for KL-3, and 251 images for KL-4.

B. Cascade-I: Label Cleaning and Adaptive Relabeling

A VGG19-based embedding network was trained using 5-fold cross-validation under a strict OOF protocol, ensuring that each sample was represented by an embedding extracted from a model that had not been trained on that sample. The network replaces the final classification layer with a 256-dimensional embedding layer. It is optimized using a triplet loss [26], where each training tuple consists of an anchor sample x_a, a positive sample x_p from the same class, and a negative sample x_n from a different class. The triplet loss is defined in (1) as follows:

$$\mathcal{L}_{triplet} = max\ (d(x_a, x_p) - d(x_a, x_n) + m, 0 \quad (1)$$

where $d(x_i, x_j) = \left\|f(x_i) - f(x_j)\right\|_2$, $f(\bullet)$ denotes the embedding function, and $m = 1.0$ ensures balanced separation between positive and negative embeddings for stable learning.

After 5-fold cross-validation, OOF embeddings are collected for all samples, and class-specific centroids are computed as:

$$\mu_c = \frac{1}{N_c}\sum_{i=1}^{N_c} e_i \quad (2)$$

where e_i represents the embedding of sample i in class c, and N_c is the number of samples in class c. A sample is flagged as suspected if its distance to the centroid exceeds a class-specific percentile threshold and its predicted label differs from the original label, with the following criterion:

$$status_i = \begin{cases} \text{suspect}, if\, d_i > T_c \text{ and } \hat{y}_i \neq y_i \\ \text{keep}, \text{otherwise} \end{cases} \quad (3)$$

where $d_i = \left||e_i - \mu_c|\right|_2$, T_c is the class-specific percentile threshold, $\hat{y}_i$ is the predicted label, and y_i is the original label. For suspect samples, ordinal-safe relabeling is performed according to (4), where y_i' is the relabeled class:

$$y_i' \in \{\max(0, y_i - 1), y_i, \min(4, y_i + 1)\} \quad (4)$$

This limits candidate labels to the original KL grade or its neighbors and removes extreme outliers above the 99th percentile. With a total of 8,260 samples, 292 were flagged, only 6 were removed, and the rest were conservatively relabeled to adjacent KL grades.

C. Cascade-II: Final Embedding and Classification

After obtaining a clean dataset from Cascade-I, the embedding network was retrained using the triplet loss defined in (1). To explicitly enforce the ordinal structure of KL grades, negative samples were selected only from classes separated by two or more grades:

$$|y_a - y_n| \geq 2 \quad (5)$$

where y_a and y_n denote the KL grades of the anchor and negative samples, respectively, as enforced in (5). This constraint increases the ordinal margin between classes, resulting in compact and well-separated embeddings that better reflect KL progression.

The optimized embeddings are then fed into a two-hidden-layer MLP classifier for KL grade prediction:

$$\hat{y} = \underset{c}{\text{argmax}}\, g_c(e) \quad (6)$$

where $e = f(x)$ is the learned embedding and $g_c(\bullet)$ denotes the output score of class c from the MLP. The classifier was trained using the predefined training, validation, and test splits summarized in Table I to ensure experimental consistency.

D. Performance Evaluation

The proposed framework was evaluated using the following metrics: accuracy, precision, recall, F1-score [27], confusion matrix, MAE [28], and QWK [29]. MAE is computed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(\hat{y}_i - y_i)| \quad (7)$$

where n is the total number of evaluated samples, $\hat{y}_i$ is the predicted KL grade, and y_i is the ground truth KL grade. Finally, QWK is calculated as :

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \quad (8)$$

where O_{ij} is the observed agreement matrix, E_{ij} is the expected agreement matrix, and w_{ij} is the quadratic weight for class pair (i, j).

IV. RESULTS AND DISCUSSION

A. Experimental Setup and Evaluation Protocol

Table II summarizes the experimental setup and evaluation protocols. A VGG19 backbone pretrained on ImageNet was employed to extract 256-dimensional embeddings, followed by a two-stage CDML pipeline comprising label reliability assessment through OOF metric learning and ordinal-aware classification to ensure consistent and reproducible evaluation.

B. Quantitative Classification Results

The proposed CDML framework was evaluated on a 20% test split (1,594 knee X-rays) covering all KL grades (0–4). The quantitative results in Table III demonstrate that CDML produces stable and balanced performance across all KL grades with no bias toward the dominant class. The error pattern, dominated by misclassification between adjacent grades, confirms the successful modeling of the ordinal structure of osteoarthritis. The integration of label reliability updating and ordinal-aware metric learning has been proven to increase

prediction consistency and reduce annotation subjectivity, making CDML more robust and relevant for clinical applications.

TABLE II. EXPERIMENTAL SETUP AND CONFIGURATION

Component	Configuration
Input image	224×224 pixels
Data split (Train:Validation:Test)	70:10:20
Backbone network	VGG19 (pretrained on ImageNet)
Embedding dimension	256
Learning rate	$1x10^{-4}$
Batch size	16
Dataset imbalance handling (Classifier)	Class weight
Cascade-I	5-fold cross-validation OOF DML, Triplet Margin Loss, AdamW, LR $1×10^{-4}$
Cascade-II	Negative sampling with $\|\Delta KL\|>2$, Triplet Margin Loss, AdamW, MLP (2 hidden layers), Weighted Binary Cross-Entropy with ordinal encoding, LR $1×10^{-3}$
Evaluation metrics	Accuracy, QWK, MAE

TABLE III. CLASSIFICATION RESULTS ON THE TEST SET

Metrics	Value
Accuracy	82.12%
Macro Precision	0.840
Macro Recall	0.853
Macro F1-score	0.846
Weighted F1-score	0.822
MAE	0.179
QWK	0.935

C. Class-Wise Performance Analysis

Table IV shows the performance of the CDML model across various KL grades. This table presents the class-wise classification results for evaluating the model robustness across individual KL grades in an ordinal, imbalanced setting. The proposed model demonstrated strong discriminative performance for KL-2, KL-3, and KL-4. In contrast, a lower performance was observed for KL-1, which can be attributed to the intrinsic ambiguity between KL-0 and KL-1. Overall, the class-wise analysis indicated improved discrimination between adjacent KL grades and confirmed the effectiveness of the ordinal-aware metric learning strategy in constraining misclassifications predominantly to neighboring KL grades rather than to non-adjacent categories.

TABLE IV. CLASS-WISE CLASSIFICATION RESULTS

KL Grade	Precision	Recall	F1-score	Accuracy (%)	Support
KL-0	0.828	0.814	0.821	81.37	655
KL-1	0.571	0.598	0.584	59.80	296
KL-2	0.945	0.926	0.936	92.60	392
KL-3	0.919	0.928	0.923	92.75	207
KL-4	0.936	1.000	0.967	100.00	44
Macro-averages	**0.840**	**0.853**	**0.846**	**85.31**	**-**

D. Confusion Matrix Analysis

To analyze test performance in detail, Figure 2 presents the confusion matrix. The CDML framework is stable, with most misclassifications occurring between adjacent KL grades and no significant error. Although KL Grade 4 was consistently predicted, some overlap existed between Grades 0 and 1 owing to visual similarities, indicating that the learned embeddings effectively encoded ordinal progression for stable, ordinal-consistent classification.

Fig. 2. Confusion Matrix

E. Ablation Study

An ablation study compared the baseline DML and CDML under identical data splits to assess the effectiveness of the latter in KL grade classification. As shown in Table V, the ablation results indicate that incorporating the full CDML pipeline leads to substantial performance improvements over the baseline DML, with an accuracy gain of 24.13% points, a 0.295 reduction in MAE, and a 0.146 increase in QWK. These gains highlight the cumulative contribution of label noise reduction, class balancing, and ordinal-aware embedding learning, demonstrating that each component plays a complementary role in improving ordinal consistency and reducing prediction error; bootstrap analysis on QWK (mean gain 0.146 with a 95% confidence interval of [0.125, 0.167]) confirms that these improvements are statistically significant.

TABLE V. PERFORMANCE COMPARISON OF BASELINE DML AND CDML ON KL GRADE OF KNEE OSTEOARTHRITIS

Model	Accuracy (%)	MAE	QWK
Baseline DML	57.99	0.474	0.789
CDML	**82.12**	**0.179**	**0.935**

To justify the choice of percentile threshold for outlier rejection, a sensitivity analysis was conducted at the 95th, 97th, and 99th percentiles, with Table VI showing their impact on the discarded samples and classification performance.

TABLE VI. PERCENTILE THRESHOLD SELECTION FOR OUTLIER REJECTION

Percentile	Drop (%)	Accuracy (%)	MAE	QWK
p95	0.47	81.99	0.181	0.936
p97	0.21	82.25	0.178	0.936
p99	0.07	82.12	0.179	0.935

As shown in Table VI, the classification performance is largely insensitive to the percentile threshold, with differences in accuracy below 0.2% and negligible variations in MAE and QWK. In contrast, the proportion of discarded samples varied substantially, with lower percentiles resulting in more aggressive data removal. Considering the minimal performance differences and the importance of retaining clinically plausible samples, the 99th percentile was selected as a conservative threshold that balanced robustness and data retention.

evaluate the influence of the CNN backbone within the CDML framework, ResNet50, EfficientNet-B0, and VGG19 were compared, and the results are summarized in Table VII. Although VGG19 is an older CNN architecture, it remains highly effective within the CDML framework. Compared with ResNet50 and EfficientNet B0, VGG19 yielded more discriminative and stable embeddings for KL grade prediction. This observation suggests that in metric-learning-based medical image analysis, particularly for knee X-ray images with subtle structural variations, embedding stability and feature consistency may be more critical than architectural complexity.

TABLE VII. PERFORMANCE COMPARISON OF CNN BACKBONES WITH CDML ON KL GRADE KNEE OSTEOARTHRITIS CLASSIFICATION

CNN backbone	Accuracy (%)	MAE	QWK
ResNet50	61.08	0.522	0.729
EfficientNet-B0	69.60	0.376	0.826
VGG19	**82.12**	**0.179**	**0.935**

Embedding robustness is further ensured through 5-fold OOF triplet learning and adaptive relabeling, which minimizes the sensitivity to noisy KL labels. The consistently low MAE and high QWK indicate that the performance gains primarily stem from the proposed CDML learning paradigm rather than the backbone architecture.

F. Grad-CAM Visualization

To illustrate the dominant features supporting the classification, Grad-CAM visualizations are presented in Table VIII, showing Grad-CAM heatmaps overlaid on the original X-ray images. KL-0 and KL-1 primarily emphasize the medial and lateral joint spaces. KL-2 and KL-3 highlight joint space narrowing and subchondral bone regions, while KL-4 focuses on bone deformities and osteophyte formation. These heatmaps consistently emphasize anatomically relevant areas associated with osteoarthritis progression.

TABLE VIII. GRAD-CAM RESULTS FOR KL GRADE CLASSIFICATION USING CDML

Image	KL-0	KL-1	KL-2	KL-3	KL-4
Original					
CDML					

G. Comparison with Previous Studies

A comparative analysis was conducted with state-of-the-art approaches reported in the literature. Table IX shows that the CDML framework outperforms previous methods, including CNN regression [15], YOLOv2 with ordinal loss [16], Deep Siamese CNN [17], ensemble networks [19], DenseNet121 [20], and CleanLab+EfficientNet B5 [21], achieving slightly higher accuracy (+0.05%) and macro F1-score (+0.027%), and notably higher QWK (+0.105%), with lower MAE, indicating better preservation of ordinal relationships. These improvements are attributed to label cleaning, adaptive relabeling, and ordinal-aware embeddings, which jointly reduce the label noise and class imbalance. Other methods, such as [18], were not compared because they relied on private data.

TABLE IX. COMPARISON WITH PREVIOUS STUDIES

Study	Dataset	Dataset size	Method	Evaluation result
[15]	OAI	8,892	CNN Regression loss	Precision = 0.610 Recall = 0.620 F1-score = 0.590
[16]	OAI	9,786	YOLOv2, CNNs with ordinal loss function	Accuracy = 69.70% MAE = 0.3440
[17]	OAI and MOST	27,293	Deep Siamese CNN	Accuracy = 66.71% QWK = 0.830 ROC = 0.930
[19]	OAI	9,786	Ensemble Networks	Accuracy = 76.93% F1-score = 0.770
[20]	OAI	9,786	Densenet121	Accuracy = 70.13% MCC = 0.5864
[21]	OAI	9,786	CleanLab+ EfficientNet B5	Accuracy = 82.07% Macro Precision = 0.805 Macro Recall = 0.803 Macro F1-score = 0.819
Proposed CDML	**OAI**	**9,786**	**CDML**	**Accuracy = 82.12% Precision = 0.840 Recall = 0.853 F1-score = 0.846 MAE = 0.179 QWK = 0.935**

V. CONCLUSION

This study presented a CDML framework for the automated KL grade classification of knee X-ray images, explicitly addressing label noise and ordinal dependencies. By integrating label reliability updating with ordinal-aware classification, CDML outperformed baseline DML on the OAI dataset, achieving an accuracy of 82.12%, MAE of 0.179, and QWK of 0.935, improving accuracy by 24.13%, reducing MAE by 0.295, and increasing QWK by 0.146. CDML with the VGG19 backbone also surpasses implementations with ResNet50 and EfficientNetB0, highlighting the effectiveness of the cascade design and ordinal-aware learning over network depth. A limitation is the absence of subject identifiers, which restricts subject-level separation across splits, and validation is limited to the OAI dataset. Future work will focus on cross-dataset evaluations and multimodal integration to further assess generalization and robustness.

REFERENCES

[1] L. Qiao *et al.*, "Epidemiological trends of osteoarthritis at the global, regional, and national levels from 1990 to 2021 and projections to 2050," *Arthritis Research & Therapy*, vol. 27, no. 1, Oct. 2025, Art. no. 199, https://doi.org/10.1186/s13075-025-03658-w.

[2] F. Muttaqin *et al.*, "A Combination Method of ROI, CLAHE, and DenseNet-169 for Hip Osteoarthritis Detection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22690–22697, June 2025, https://doi.org/10.48084/etasr.10576.

[3] D. J. Hunter, L. March, and M. Chew, "Osteoarthritis in 2020 and beyond: a Lancet Commission," *The Lancet*, vol. 396, no. 10264, pp. 1711–1712, Nov. 2020, https://doi.org/10.1016/S0140-6736(20)32230-3.

[4] K. S. Basavaraju, T. K. Kumar, and K. A. Reddy, "Vibroarthrographic Signal Classification for Knee Joint Disorder Detection using Tunable Q-factor Wavelet Transform based on Entropy Measures," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19953–19958, Feb. 2025, https://doi.org/10.48084/etasr.9245.

[5] J. H. Kellgren and J. S. Lawrence, "Radiological Assessment of Osteo-Arthrosis," *Annals of the Rheumatic Diseases*, vol. 16, no. 4, pp. 494–502, Dec. 1957, https://doi.org/10.1136/ard.16.4.494.

[6] J. S. Yoon *et al.*, "Assessment of a novel deep learning-based software developed for automatic feature extraction and grading of radiographic knee osteoarthritis," *BMC Musculoskeletal Disorders*, vol. 24, no. 1, Nov. 2023, Art. no. 869, https://doi.org/10.1186/s12891-023-06951-4.

[7] S. Beyaz, S. B. Yayli, and K. Kılıç, "From variability to consistency: building a Kellgren-Lawrence gonarthrosis dataset," *Journal of Orthopaedic Surgery and Research*, vol. 20, no. 1, Oct. 2025, Art. no. 922, https://doi.org/10.1186/s13018-025-06057-8.

[8] A. G. Culvenor, C. N. Engen, B. E. Øiestad, L. Engebretsen, and M. A. Risberg, "Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 23, no. 12, pp. 3532–3539, Dec. 2015, https://doi.org/10.1007/s00167-014-3205-0.

[9] K. Klara *et al.*, "Reliability and Accuracy of Cross-sectional Radiographic Assessment of Severe Knee Osteoarthritis: Role of Training and Experience," *The Journal of Rheumatology*, vol. 43, no. 7, pp. 1421–1426, July 2016, https://doi.org/10.3899/jrheum.151300.

[10] A. S. Buriboev, A. Abduvaitov, and H. S. Jeon, "Binary Classification of Pneumonia in Chest X-Ray Images Using Modified Contrast-Limited Adaptive Histogram Equalization Algorithm," *Sensors*, vol. 25, no. 13, June 2025, Art. no. 3976, https://doi.org/10.3390/s25133976.

[11] V. Jain, A. K. Dubey, and A. Jain, "SCDNet 1.0: Adaptive CNN Framework for Sickle Cell Disease Detection with OTSU Segmentation and Gaussian Filter," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, Dec. 2025, Art. no. 18758967251405463, https://doi.org/10.1177/18758967251405463.

[12] R. R. Sarra, A. E. Korial, I. I. Gorial, and A. J. Humaidi, "Enhancing Migraine Classification Through Machine Learning: A Comparative Study of Ensemble Methods," *Technologies*, vol. 13, no. 11, Nov. 2025, Art. no. 500, https://doi.org/10.3390/technologies13110500.

[13] Y. Wang, X. Wang, T. Gao, L. Du, and W. Liu, "An Automatic Knee Osteoarthritis Diagnosis Method Based on Deep Learning: Data from the Osteoarthritis Initiative," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–10, Sept. 2021, https://doi.org/10.1155/2021/5586529.

[14] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, Oct. 2020, Art. no. 101759, https://doi.org/10.1016/j.media.2020.101759.

[15] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 1195–1200, https://doi.org/10.1109/ICPR.2016.7899799.

[16] 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, July 2019, https://doi.org/10.1016/j.compmedimag.2019.06.002.

[17] 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, Jan. 2018, Art. no. 1727, https://doi.org/10.1038/s41598-018-20132-7.

[18] J. Pan *et al.*, "Automatic knee osteoarthritis severity grading based on X-ray images using a hierarchical classification method," *Arthritis Research & Therapy*, vol. 26, no. 1, Nov. 2024, Art. no. 203, https://doi.org/10.1186/s13075-024-03416-4.

[19] S. W. Pi, B. D. Lee, M. S. Lee, and H. J. Lee, "Ensemble deep-learning networks for automated osteoarthritis grading in knee X-ray images," *Scientific Reports*, vol. 13, no. 1, Dec. 2023, Art. no. 22887, https://doi.org/10.1038/s41598-023-50210-4.

[20] Y. Wang *et al.*, "Learning From Highly Confident Samples for Automatic Knee Osteoarthritis Severity Assessment: Data From the Osteoarthritis Initiative," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1239–1250, Mar. 2022, https://doi.org/10.1109/JBHI.2021.3102090.

[21] T. Momenpour and A. Abu Mallouh, "Optimizing CNN-Based Diagnosis of Knee Osteoarthritis: Enhancing Model Accuracy with CleanLab Relabeling," *Diagnostics*, vol. 15, no. 11, May 2025, Art. no. 1332, https://doi.org/10.3390/diagnostics15111332.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015, https://doi.org/10.48550/arXiv.1409.1556.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," presented at the Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017, Art. no. 618–626.

[24] P. Chen, "Knee Osteoarthritis Severity Grading Dataset." Mendeley, Sept. 04, 2018, https://doi.org/10.17632/56RMX5BJCR.1.

[25] J. B. Driban *et al.*, "The state of the Osteoarthritis Initiative (OAI): Entering a new era," *Seminars in Arthritis and Rheumatism*, vol. 75, Dec. 2025, Art. no. 152887, https://doi.org/10.1016/j.semarthrit.2025.152887.

[26] E. Hoffer and N. Ailon, "Deep Metric Learning Using Triplet Network," in *Similarity-Based Pattern Recognition*, vol. 9370, A. Feragen, M. Pelillo, and M. Loog, Springer International Publishing, 2015, pp. 84–92.

[27] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv, Oct. 11, 2020, https://doi.org/10.48550/arXiv.2010.16061.

[28] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, pp. 79–82, 2005, https://doi.org/10.3354/cr030079.

[29] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968, https://doi.org/10.1037/h0026256.