ORIGINAL ARTICLE

COMPARISON OF THE PERFORMANCE OF PRETRAINED DEEP LEARNING MODELS FOR THE AUTOMATIC KELLGREN-LAWRENCE GRADING OF KNEE OSTEOARTHRITIS USING PLAIN RADIOGRAPHS

Radyografiler Kullanılarak Diz Osteoartritinin Otomatik Kellgren-Lawrence Derecelendirilmesi için Önceden Eğitilmiş Derin Öğrenme Modellerinin Karşılaştırmalı Değerlendirmesi

Hafize KIZILKAYA[1], Fatma Nur ORTATAS[2], Kemal URETEN[1]

[1]Bozok Üniversitesi,
Tıp Fakültesi,
İç Hastalıkları Anabilim Dalı,
Yozgat,
Türkiye.
[2]Bozok Üniversitesi,
Mühendislik-Mimarlık Fakültesi,
Bilgisayar Mühendisliği Bölümü,
Yozgat,
Türkiye.

Hafize KIZILKAYA, Dr. Öğr. Ü.
(https://orcid.org/0000-0002-4878-9958)
✉ hfz.kizilkaya@gmail.com
Fatma Nur ORTATAŞ, Bilg. Müh.
(https://orcid.org/0000-0001-7897-9958)
✉ f.nur.ortatas@yobu.edu.tr
Kemal ÜRETEN, Prof. Dr.
(https://orcid.org/0000-0002-7673-4399)
✉ kemalureten@yahoo.com

İletişim:
Dr. Öğr. Ü. Hafize KIZILKAYA
Bozok Üniversitesi, Tıp Fakültesi, İç Hastalıkları Anabilim Dalı.
Merkez, Yozgat, Türkiye.

Geliş tarihi/Received: 08.01.2026
Kabul tarihi/Accepted: 24.02.2026
DOI: 10.16919/bozoktip.1859321

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

ABSTRACT

Objective: This study aimed to comparatively evaluate pre-trained deep learning architectures for automated Kellgren-Lawrence (KL) grading of knee osteoarthritis from plain radiographs within an ordinal-aware learning framework.

Material and Methods: A total of 8260 knee radiographs labeled with KL grades 0-4 from the Osteoarthritis Initiative (OAI) dataset were retrospectively analyzed. The dataset included predefined training (70%), validation (10%), and independent test (20%) partitions. Five pre-trained convolutional neural network backbones (VGG-16, ResNet-50, DenseNet-121, EfficientNetB0, and InceptionV3) were adapted via transfer learning under identical training conditions to ensure controlled comparison. To explicitly model the ordered structure of KL grades, an ordinal CORAL-based loss function was employed. Model selection was performed using five-fold stratified cross-validation, with Quadratic Weighted Kappa (QWK) as the primary evaluation metric. Secondary metrics included overall accuracy, balanced accuracy, macro-F1 score, class-wise precision and recall, confusion matrix analysis, receiver operating characteristic (ROC), and precision-recall (PR) analyses. Decision curve analysis (DCA) was conducted at clinically relevant thresholds (KL ≥2 and KL ≥3). Final evaluation was performed on the independent test set.

Results: All architectures demonstrated comparable ordinal agreement. VGG-16 achieved the highest test-set performance (QWK =0.830, macro-F1=0.676, balanced accuracy =0.684). Discriminative performance was stronger in moderate-to-severe stages (KL 3-4), whereas reduced sensitivity was observed in early-stage disease, particularly KL Grade 1. Misclassifications predominantly occurred between adjacent grades. ROC, PR, and decision curve analyses supported clinical utility.

Conclusion: Ordinal-aware deep learning provides reliable automated KL grading from plain radiographs. External multi-center validation is required to enhance generalizability.

Keywords: *Kellgren-Lawrence Grading; Knee Osteoarthritis; Deep Learning; Ordinal Classification; Quadratic Weighted Kappa; Decision Curve Analysis*

ÖZET

Amaç: Bu çalışma, diz osteoartritinin Kellgren-Lawrence (KL) derecelendirmesinde önceden eğitilmiş derin öğrenme mimarilerini, düz radyografiler üzerinden ordinal-farkındalıklı bir çerçevede karşılaştırmayı amaçlamaktadır.

Gereç ve Yöntemler: Osteoarthritis Initiative (OAI) veri setinde yer alan ve KL dereceleri 0-4 arasında etiketlenmiş 8260 diz radyografisi retrospektif olarak analiz edilmiştir. Veri seti önceden tanımlanmış eğitim (%70), doğrulama (%10) ve bağımsız test (%20) bölümlerinden oluşmaktadır. Beş önceden eğitilmiş evrişimsel sinir ağı mimarisi (VGG-16, ResNet-50, DenseNet-121, EfficientNetB0 ve InceptionV3) transfer öğrenme ile aynı eğitim koşulları altında uyarlanmıştır. KL derecelerinin sıralı yapısını modellemek amacıyla ordinal CORAL tabanlı kayıp fonksiyonu kullanılmıştır. Model seçimi beş katlı stratifiye çapraz doğrulama ile gerçekleştirilmiş ve birincil değerlendirme metriği olarak Quadratic Weighted Kappa (QWK) kullanılmıştır. İkincil ölçütler arasında doğruluk, dengeli doğruluk, makro-F1, sınıf bazlı duyarlılık ve kesinlik, karışıklık matrisi, ROC ve precision-recall analizleri yer almıştır. Klinik eşiklerde (KL ≥2 ve KL ≥3) karar eğrisi analizi uygulanmıştır. Nihai değerlendirme bağımsız test setinde yapılmıştır.

Bulgular: Tüm mimariler benzer ordinal uyum göstermiştir. VGG-16 test setinde en yüksek QWK (0,830), makro-F1 (0,676) ve dengeli doğruluk (0,684) değerlerine ulaşmıştır. Orta ve ileri evrelerde (KL 3-4) performans daha yüksek, özellikle KL Evre 1'de duyarlılık daha düşüktür. Yanlış sınıflandırmalar çoğunlukla komşu dereceler arasında gerçekleşmiştir. ROC, PR ve karar eğrisi analizleri klinik faydayı desteklemiştir.

Sonuç: Ordinal-farkındalıklı derin öğrenme yaklaşımı, diz osteoartritinin otomatik KL derecelendirilmesinde güvenilir sonuçlar sunmaktadır. Klinik genellenebilirlik için çok merkezli dış doğrulama gereklidir.

Anahtar Kelimeler: *Kellgren-Lawrence Derecelendirme; Diz Osteoartriti; Derin Öğrenme; Ordinal Sınıflandırma; Quadratic Weighted Kappa; Karar Eğrisi Analizi*

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

INTRODUCTION

Osteoarthritis (OA) is a degenerative joint disease that is ranked among the frequently occurring degenerative diseases in the globe, which causes a lot of pain, loss of functional ability, and reduced quality of life, especially in the elderly. Knee osteoarthritis has become a significant issue of public health because its prevalence is high and it severely affects the day-to-day lives of people. The clinical decision-making processes, disease surveillance and treatment planning require early diagnosis and precise determination of the severity of the disease (1). Despite OA being previously believed to be a degenerative disease with progressive loss of cartilages in the joint, it is currently appreciated that it is a complex disease, which involves subchondral bone remodeling, osteophyte formation, synovial inflammation, and alteration in the periarticular tissues besides cartilage degeneration (2,3).

The pathogenesis of osteoarthritis is a multifactorial one and the cause of the disease is the combination of mechanical loading, ageing, genetic predisposition, metabolic factors, and low-grade chronic inflammation, which lead to the gradual destruction of the joint structure. According to epidemiological data, it has been noted that OA is very prevalent in the ageing population and that knee osteoarthritis in particular is one of the leading causes of functional impairment and health care burden because it is the joint that is most affected. Thus, as the disease is progressive, and it has high impacts on the society, it is important that it should be staged and diagnosed early to be able to enhance management of the disease and its outcomes on patients (3).

Radiographic assessment plays a central role in the clinical evaluation of knee osteoarthritis, and the Kellgren-Lawrence (KL) grading system is the most commonly used method for classifying disease severity. First defined in 1957, the KL system classifies knee osteoarthritis on a scale from 0 (no radiographic OA findings) to 4 (advanced OA) based on the presence and degree of osteophyte formation, joint space narrowing, subchondral sclerosis, and bone deformities (4,5). The Kellgren-Lawrence (KL) grading system is presented in Table 1.

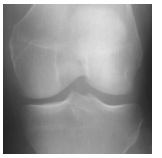
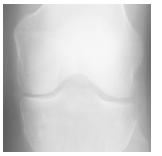
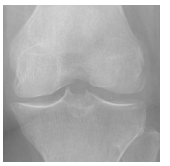
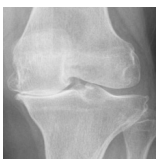
Due to its simplicity, low cost, and widespread acceptance in clinical and epidemiological studies, the KL grading system continues to be used as a fundamental tool in the diagnosis and research of knee osteoarthritis (OA)(6). Nevertheless, the KL grading system has well-defined limitations. The subjective nature of the grading criteria leads to significant inter-observer and intra-observer variability, particularly in early-stage OA where radiographic changes are unclear (7). This variability reduces both clinical reliability and reproducibility, especially in distinguishing subtle differences between neighboring grades (e.g., KL1 vs. KL2). Additionally, the Kellgren-Lawrence (KL) grading system provides a relatively crude representation of disease severity and may not fully reflect the heterogeneous phenotypes of osteoarthritis (OA), symptom variations, or the inconsistencies between radiographic severity and clinical pain/functional limitations (8). These limitations emphasize the necessity of more objective, quantitative approaches to assessing knee OA and encourage the development of automated imaging techniques that aim to reduce subjectivity and improve classification consistency.

Recent advances in medical image analysis have highlighted the use of deep learning (DL) techniques to increase objectivity and reproducibility in the assessment of musculoskeletal disorders. Convolutional neural networks (CNNs), in particular, have demonstrated successful results in tasks such as disease detection and severity grading on radiographic images. In the field of knee osteoarthritis, DL-based approaches are increasingly being researched with the aim of automating Kellgren-Lawrence (KL) grading from plain radiographs and reducing subjectivity dependent on human interpretation.

In recent years, numerous studies utilizing pre-trained CNN architectures have reported high classification performance in predicting KL grades from knee radiographs (9-12). These studies demonstrate that transfer learning-based approaches can effectively learn radiographic features such as osteophyte formation and joint space narrowing. However, it has been noted that the reported accuracy and area under the curve (AUC) values vary significantly depending on the dataset size, class distribution, and classification strategies (13,17). Particularly in five-class KL grading, it has been reported that classification performance decreases and clinical generalizability remains limited due to subtle

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

Table 1. Kellgren-Lawrence (KL) grading system for knee osteoarthritis

KL grade	Osteoarthritis severity	Representative radiograph	Radiographic characteristics
Grade 0	Normal		No radiographic features of osteoarthritis.
Grade 1	Doubtful OA		Doubtful joint space narrowing and possible osteophyte formation.
Grade 2	Mild OA		Possible joint space narrowing with definite osteophyte formation.
Grade 3	Moderate OA		Multiple osteophytes, definite joint space narrowing, and subchondral sclerosis.
Grade 4	Severe OA		Marked joint space narrowing, large osteophytes, severe subchondral sclerosis, and definite bone deformity.

OA: Osteoarthritis

differences between neighboring stages (13,14,16). Therefore, there remains a need for advanced approaches to enhance both discriminative power and clinical reliability in multi-class KL grading.

In contrast to many previous studies that formulate Kellgren-Lawrence (KL) grading as a purely nominal multi-class classification task, the present study explicitly models the ordinal nature of disease severity using an ordinal-aware learning framework and evaluation strategy.

First, an ordinal CORAL-based loss function combined with quadratic weighted kappa (QWK) was adopted to better reflect the progressive structure of KL grades and to penalize clinically implausible misclassifications. Second, multiple widely used pre-trained convolutional neural network architectures were compared under strictly identical training, augmentation, and class-balancing conditions, enabling a controlled and fair assessment of backbone-specific behavior rather than architecture-dependent optimization. Third, beyond conventional performance metrics, the clinical utility of the proposed model was evaluated using decision curve analysis at clinically relevant thresholds (KL ≥ 2 and KL ≥ 3), providing insight into potential decision-support value across both early and advanced stages of knee osteoarthritis.

Together, these contributions aim to improve the interpretability, clinical relevance, and methodological rigor of automated KL grading from plain radiographs.

Accordingly, this study aims to comparatively evaluate pre-trained deep learning models for automated Kellgren-Lawrence grading of knee osteoarthritis from plain radiographs within an ordinal-aware learning and evaluation framework.

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

MATERIAL AND METHOD

This retrospective experimental study aimed to comparatively evaluate the performance of different pre-trained deep learning models for the automatic Kellgren-Lawrence grading of knee osteoarthritis using plain radiographs. The study was approved by the local ethics committee (Date: 29.12.2025; Approval No:867).

The experiments were conducted using knee radiographs obtained from the publicly available Osteoarthritis Initiative (OAI) dataset (18). The OAI provides large, standardized radiographic images with Kellgren-Lawrence (KL) grades ranging from 0 to 4, assigned by experts, designed to investigate the development and progression of knee osteoarthritis. A total of 8,260 knee radiographic images were included in this study. The dataset provides predefined training, validation, and test partitions, corresponding to a 70/10/20 split.

The predefined splits were not regenerated manually. Inspection of the accompanying metadata confirmed that each subject appears as a single record, with left and right knee radiographs associated with the same subject entry, and that subject records are not distributed across different partitions.

For model development and selection, the original training and validation subsets were merged and jointly used for 5-fold stratified cross-validation. This cross-validation was employed exclusively for model selection and robustness assessment. The predefined test set was kept fully independent and was used only once for final performance evaluation.

The input to the models consisted of knee radiograph images, each containing a single knee joint. No bilateral radiographs were included. As a result, the entire image inherently corresponds to the knee region of interest, and no additional ROI extraction was performed. All images were used at 224×224 input resolution; when required by the backbone, images were resized to match the network input size. No ROI cropping was applied. Therefore, no additional resizing or cropping operations were applied. To enhance model generalization and address class imbalance, data augmentation was applied exclusively to the training data using an online stochastic approach. Augmentation operations were restricted to clinically plausible transformations, including small-angle rotations (±7°) to simulate minor patient positioning variations, as well as slight adjustments in brightness and contrast to model acquisition variability. No histogram equalization or contrast enhancement techniques, such as contrast limited adaptive histogram equalization (CLAHE), were applied during preprocessing.

This study employed five commonly used pre-trained convolutional neural network architectures (VGG-16, ResNet-50, DenseNet-121, EfficientNetB0 and InceptionV3) for a comparative performance evaluation. All models were pre-trained on the ImageNet dataset and adapted for knee osteoarthritis classification using transfer learning. To ensure a fair comparison, all models were trained using the same evaluation parameters.

To ensure a consistent and fair comparison across architectures, all models were trained under identical experimental settings, including the same input resolution, optimization strategy, and stopping criteria. Hyperparameters were determined based on preliminary experiments and aligned with commonly adopted practices in medical image classification. During training, online data augmentation was combined with class-balanced sampling strategies to mitigate class imbalance and to ensure adequate representation of minority classes, particularly Kellgren-Lawrence Grade 4, across training iterations. Detailed training configurations and hyperparameter settings are provided in Supplementary Table 1.

All backbone architectures were trained using identical optimization strategies and hyperparameter configurations to enable a controlled comparison under the same training conditions, rather than to maximize individual model performance. Due to the pronounced class imbalance in the dataset, preliminary experiments explored both sampling-based and loss-based balancing strategies. As class weighting alone provided limited benefit when combined with the ordinal loss formulation, a WeightedRandomSampler was employed during training to equalize sampling probabilities across classes while preserving ordinal structure.

To account for the ordered nature of Kellgren-Lawrence grades, the standard CORAL loss formulation was adopted. The model output consisted of $K-1=4$ ordinal thresholds corresponding to cumulative

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

thresholds corresponding to cumulative probabilities P(KL≥k) for k=1,…,4. These cumulative outputs were subsequently used for downstream analyses, including receiver operating characteristic (ROC), precision-recall (PR), and decision curve analysis (DCA), by evaluating clinically relevant thresholds such as KL ≥ 2 and KL ≥3. In practice, CORAL threshold outputs were converted to a final KL grade (0-4) for reporting class-wise metrics.

The models have been utilized to compare the overall classification performance metrics with the ordinal characteristics of the Kellgren-Lawrence (KL) grading scale. The Quadratic Weighted Kappa (QWK) was employed as the primary evaluation metric since it addresses non-agreement discrepancies on a KL scale and is frequently utilized for assessing agreement in ordinal classification assignments. The metrics for secondary examination included overall accuracy, balanced accuracy, and macro-averaged F1 score. The precision, recall (sensitivity), and F1-score for each class were presented to evaluate performance across specific KL grades, with particular emphasis on the minority class (Grade 4). Confusion matrices, in both raw and normalized forms, were utilized to analyze model behavior. Receiver operating characteristic (ROC) curves and corresponding area under the curve (AUC) for each class were computed using a one-vs-rest (OvR) approach. Due to the intrinsic class imbalance of the dataset, precision-recall (PR) curves were constructed with average precision (AP) scores. In the cross-validation experiment, all performance metrics were reported as mean and standard deviation across folds. The QWK validation was conducted to facilitate early stopping and to select a model that mitigates overfitting while ensuring robust model selection.

Decision curve analysis was performed to assess the clinical utility of the proposed ordinal deep learning model. Although the model was trained for five-class Kellgren-Lawrence grading, DCA was applied by converting predictions into clinically relevant binary decision scenarios (KL ≥ 2 vs. KL < 2 and KL ≥ 3 vs. KL < 3). Individual-level predicted probabilities were derived from CORAL-based ordinal outputs. Net benefit was calculated across a range of threshold probabilities and compared with default strategies of treating all patients or treating none.

RESULTS

The performance of the proposed framework was evaluated across multiple backbone architectures using five-fold stratified cross-validation. Model performance was primarily assessed using quadratic weighted kappa (QWK), with secondary metrics including accuracy, balanced accuracy and macro-averaged F1-score providing additional insight. Additionally, class-wise precision, recall and F1-score were computed for each Kellgren-Lawrence (KL) grade. Confusion matrices, receiver operating characteristic (ROC) curves and precision-recall (PR) curves were then used to analyze the models' behavior further. For all cross-validation experiments, results are reported as the mean±standard deviation across folds.

On the independent test set, VGG16 achieved a Quadratic Weighted Kappa (QWK) of 0.830, a macro-F1 score of 0.676, and a balanced accuracy of 0.684, indicating robust ordinal agreement across KL grades. In addition to these global metrics, class-wise performance-including recall (sensitivity), precision, and F1-score is systematically summarized in Table 2.

Although VGG16 achieved slightly higher test-set performance, the overall differences between top-performing backbones were modest. In particular, VGG16 and EfficientNetB0 demonstrated comparable ordinal agreement, suggesting that multiple

Table 2. Five-fold stratified cross-validation performance of the evaluated backbone architectures

Backbone	QWK (mean ± std)	Accuracy (mean)	Balanced Acc. (mean ± std)	Macro-Precision	Macro-Recall / Sensitivity	Macro-F1 (mean ± std)
DenseNet121	0.813±0.006	0.6124	0.660±0.003	0.687	0.653	0.649±0.015
EfficientNetB0	0.819±0.003	0.6250	0.670±0.013	0.704	0.685	0.669±0.013
InceptionV3	0.807±0.002	0.6223	0.661±0.017	0.670	0.670	0.655±0.008
ResNet50	0.812±0.007	0.6238	0.657±0.019	0.689	0.665	0.654±0.015
VGG16	0.819±0.011	0.6477	0.675±0.009	0.699	0.665	0.664±0.012

QWK: quadritic weighted kappa

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

architectures can effectively capture radiographic patterns relevant to KL grading under identical training conditions. Class-wise evaluation revealed heterogeneous performance across KL grades. Higher precision and recall values were generally observed for KL Grade 0 and Grade 2, reflecting clearer radiographic patterns at these stages. In contrast, KL Grade 1 and Grade 4 exhibited lower sensitivity, with Grade 4 showing the greatest variability due to severe class imbalance. Despite the application of class balancing strategies, misclassification in higher grades remained more frequent, indicating the intrinsic difficulty of modeling rare and advanced disease stages.

Confusion matrix analysis further illustrated model behavior across neighboring KL grades. Confusion matrix analysis showed that misclassifications predominantly occurred between adjacent KL grades (particularly KL 1-2 and KL 2-3), while distant grade confusions were rare, with most KL Grade 4 errors occurring as misclassification into KL Grade 3, consistent with clinical adjacency (Figure 1(a)). Receiver operating characteristic (ROC) and precision-recall (PR) curve analyses were performed in a one-vs-rest (OvR) manner to assess class-wise discriminative performance of the VGG16-based model on the test set (Figure 1(b)).

Performance varied across Kellgren-Lawrence (KL) grades on the independent test set, with higher discriminative performance observed for normal and advanced disease stages. Receiver operating characteristic (ROC) analysis demonstrated strong separability for KL grade 3 (AUC=0.956) and KL grade 4 (AUC=0.946), indicating robust discrimination for advanced osteoarthritis. Similarly, KL grade 0 achieved a high AUC value of 0.908, suggesting reliable identification of normal joints. In contrast, KL grade 1 exhibited comparatively lower discriminative performance (AUC=0.747), reflecting the intrinsic difficulty of distinguishing early-stage osteoarthritic changes from normal radiographic appearances. KL grade 2 achieved a moderate performance with an AUC of 0.869.

Precision-recall (PR) analysis revealed a consistent trend. High average precision (AP) values were obtained for KL grades 0 (AP=0.845), 3 (AP=0.825), and 4 (AP=0.816), whereas KL grade 2 yielded a moderate AP of 0.715. Notably, KL grade 1 demonstrated substantially lower precision-recall performance (AP=0.337), indicating pronounced class overlap and reduced precision, particularly at higher recall levels (Figure 1(c)). Overall, these findings suggest that the proposed model effectively differentiates between normal and advanced osteoarthritis stages, while classification of early and borderline grades-especially KL grade 1-remains challenging, consistent with the inherent ambiguity of the KL grading system.

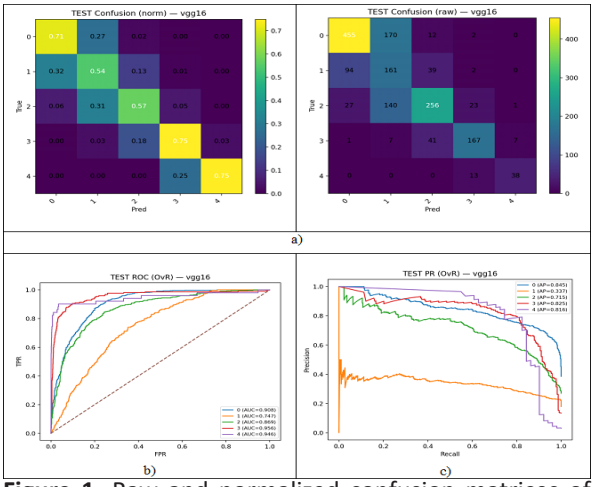


Figure 1. Raw and normalized confusion matrices of the VGG-16 model on the independent test set(a). Receiver operating characteristic (ROC) curves (b) and precision-recall (PR) curves (c) for class-wise evaluation of the VGG16-based model on the test set

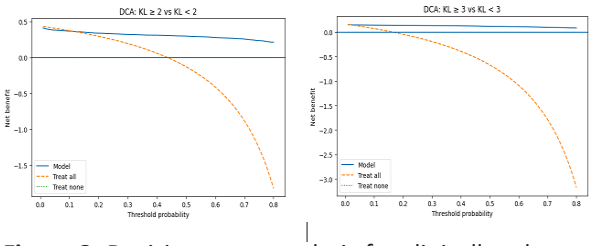


Figure 2. Decision curve analysis for clinically relevant decision thresholds. (a) KL ≥2 vs. KL <2 and (b) KL ≥3 vs. KL < 3. The proposed model is compared with default strategies of treating all patients and treating none across a range of threshold probabilities

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

Decision curve analysis demonstrated that the proposed ordinal deep learning model provided a consistently higher net benefit than both treat-all and treat-none strategies across a wide range of clinically relevant threshold probabilities for both decision scenarios (Figure 2). In the KL ≥2 versus KL <2 scenario, corresponding to the identification of definite radiographic osteoarthritis, the model achieved positive net benefit particularly at low-to-moderate threshold probabilities, indicating its utility for early-stage decision-making and follow-up planning. In the KL ≥3 versus KL <3 scenario, representing moderate-to-severe osteoarthritis, the model maintained a positive net benefit across an even broader range of threshold probabilities, highlighting its potential role in supporting decisions related to intensified management and advanced interventions. Together, these findings suggest that the proposed model may offer meaningful clinical utility across both early and advanced stages of knee osteoarthritis.

The inclusion of both KL ≥2 and KL ≥3 decision thresholds was motivated by their distinct clinical implications. The KL ≥2 threshold reflects the presence of definite radiographic osteoarthritis and is primarily associated with early clinical decisions such as monitoring, lifestyle modification, and initiation of conservative treatment. In contrast, the KL ≥3 threshold represents moderate-to-severe disease, for which decisions regarding intensified treatment strategies, advanced imaging, or surgical consultation may be considered. By evaluating decision curve analysis at both thresholds, the present study demonstrates that the proposed model provides clinical utility not only for identifying advanced disease but also for supporting early-stage decision-making. This dual evaluation highlights the model's robustness across different stages of disease severity and varying levels of clinical risk tolerance.

DISCUSSION

This paper shows that consistent and clinically plausible results with automated knee radiograph grading with ordinal-aware deep learning models can be obtained on plain knee radiographs. The findings show that deep learning-based methods are capable of learning radiographic patterns of plain knee radiographs and can maintain consistent multi-class performance, especially when ordinal-sensitive evaluation measures like quadratic weighted kappa (QWK) are used. These results are consistent with other researchers that suggested that transfer learning is appropriate in automated KL grading. In all the assessed architectures, moderate-to-severe stages of osteoarthritis (KL Grades 3 and 4) displayed a higher level of discriminative performance as the ROC-AUC and average precision values showed. It follows this tendency as radiographic appearances of more advanced disease stages generally have more striking appearances with a more narrow joint space, an osteophyte, and subchondral sclerosis

Table 3. Summary of recent deep learning-based studies for automated Kellgren-Lawrence grading, highlighting classification setup and reported evaluation metrics

Reference	Year	KL Classes	Task formulation	Reported metrics
(17)	2023	3	Multi-class	Accuracy, Recall, F1
(19)	2022	5	Nominal	Accuracy, Recall, F1
(20)	2020	5	Nominal	Accuracy, Recall, F1
(21)	2020	5	Nominal	Accuracy, Loss
(22)	2021	5	Nominal / Ordinal	Accuracy, MAE, QWK
(23)	2022	5	Nominal	Accuracy, Recall, F1
(24)	2021	5	Nominal / Ordinal	Accuracy, MAE, QWK
(25)	2022	5	Nominal	Accuracy, Precision, F1
(26)	2022	5	Nominal	Accuracy
(27)	2022	2	Binary	Accuracy, Loss
(28)	2023	2	Binary	Accuracy, Recall, F1
This study	2025	5	Ordinal-aware	QWK, Balanced Acc., Macro-F1, ROC, PR, DCA

KL: Kellgren-Lawrence, MAE: Mean absolute error, QWK: Quadratic weighted Kappa, ROC: Receiver Operating Characteristic, PR: Precision-Recall, DCA: Decision curve analysis

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

which makes it easier to extract features automatically. KL Grade 1 consistently exhibited lower performance across multiple evaluation metrics, reflecting its well-documented borderline nature and the inherent difficulty of distinguishing early osteoarthritic changes from normal radiographic appearances. This finding is consistent with previous reports highlighting both inter-observer variability and limited radiographic conspicuity at early disease stages.

Comparing with the already published research that has been summarized in, it becomes apparent that there is a significant difference in the methodological design and the interpretation of the performance (Table 3) (17,19-28). Most of the previous studies present high overall accuracy rates of five-class KL grading, but in most cases, those studies develop the task as a nominal multi-class classification problem and use global measures like accuracy, recall, or F1-score. These measures can be highly dependent on the class imbalance and they do not necessarily punish clinically implausible misclassifications between distantly related KL grades. Conversely, the current analysis will use an ordinal-conscious learning model and make quadratic weighted kappa (QWK) the main evaluation measure that will allow a more clinically valuable evaluation of consensus between ordered stages of the disease. Even though the general accuracy obtained in this study is relatively less than certain values reported in the literature, the delivered QWK falls within the domain of the most powerful previously published ordinal-based methods. In addition, the analysis of errors shows that false classifications mainly happen between neighboring KL grades, which justify the clinical plausibility and interpretability of the suggested framework beyond the comparisons of raw accuracy.

Confusion matrix analysis also showed that confusion of one grade with its neighboring grade was predominant with few instances of confusion between two distant grades. The predominance of misclassifications between adjacent KL grades further supports the clinical plausibility of the observed error patterns and motivates a qualitative investigation of the model's

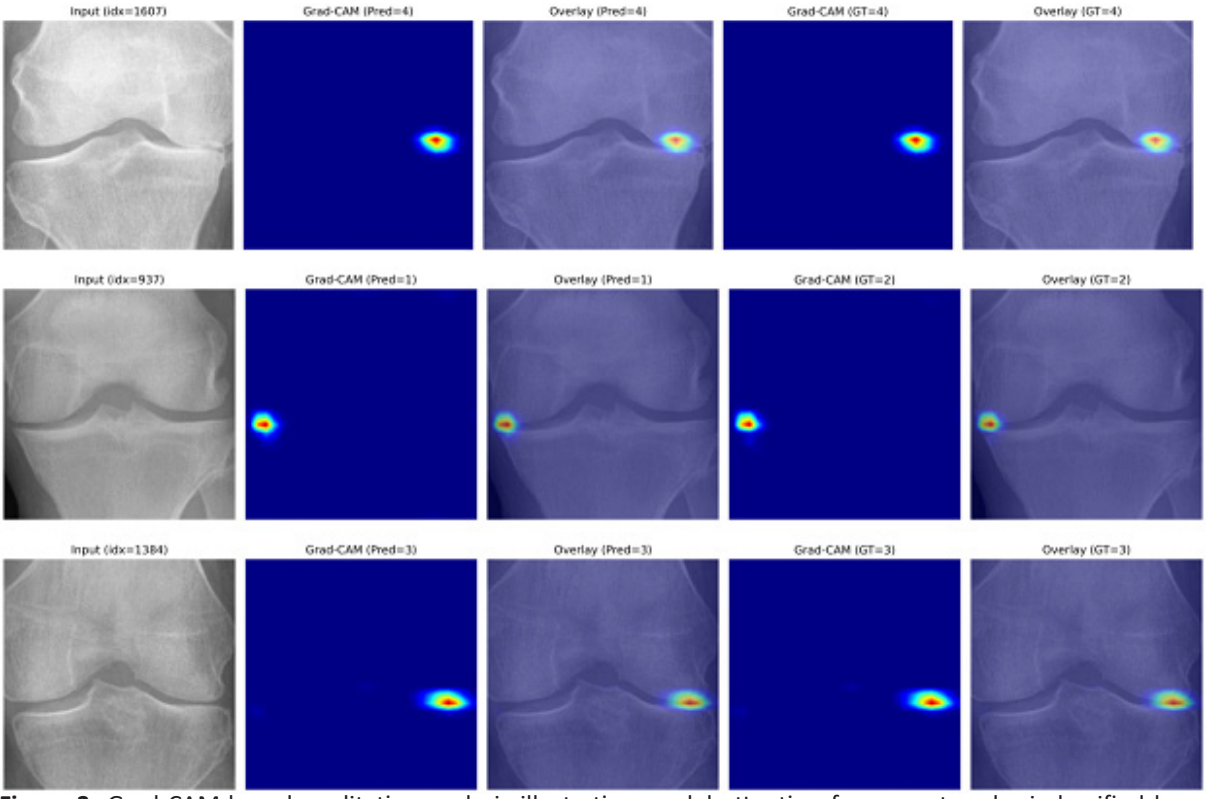


Figure 3. Grad-CAM-based qualitative analysis illustrating model attention for correct and misclassified knee radiographs across Kellgren-Lawrence grades

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

decision-making process. It is worth noting that out of the considered backbone architectures, VGG-16 proved to be a little more stable and slightly better performing than the deeper or more complicated ones. This finding indicates that the simplicity of architectural features and conservation of global geometric aspects can prove beneficial to radiographic osteoarthritis grading, and discriminative patterns are in many cases faint and widely spread.

It must be noted that there are a number of limitations of this study. A single publicly available dataset was used to conduct the experiments and this might not be applicable in generalizing the results to data obtained at varying imaging conditions or basing the results in different populations. Moreover, only radiographic images were used in the analysis and no clinical or demographic data among participants was included, and it could affect osteoarthritis severity. Moreover, they did not tune hyperparameters with regards to the backbone and did not perform explicit subject-level re-splitting since the research was supposed to be a controlled comparative evaluation where the same training conditions were used.

To learn more about how the model makes decisions and to look at patterns of wrong classification, Gradient-weighted Class Activation Mapping (Grad-CAM) images were made of representative radiographs that were successfully and incorrectly classified across all Kellgren-Lawrence grades. As shown in Figure 3, the activation maps repeatedly showed clinically important anatomical areas, mainly the tibiofemoral joint space and areas of marginal bone that are often linked to narrowing of the joint space and the formation of osteophytes.

When the right classification was made, especially for moderate to severe osteoarthritis (KL Grades 3 and 4), Grad-CAM maps showed that areas with severe structural degeneration had activations that were compact and consistent in space. Early and borderline stages, like KL Grade 1, on the other hand, had activations that were more localized but weaker and less spread out. This was because the radiography features in these stages are often subtle and unclear.

Misclassifications mostly happened between KL grades that were next to each other, and they were accompanied by overlapping activation patterns in anatomically similar regions instead of focus on areas of the image that weren't important. This finding shows that mistakes in classification are caused by the fact that x-rays can't always tell the difference between nearby disease stages, not by models that aren't stable or false feature learning. These qualitative results are in line with the confusion matrix and ordinal performance analyses. They also add to the clinical support for the suggested ordinal-aware framework.

Areas of future work should include external validation with multi-center data sets, combination with other clinical data and study of alternative grading or regression-based methods to overcome the inherent constraints of the KL system especially at the early disease stage. On the whole, the results can be summarized as that, automated KL grading based on plain radiographs with the help of deep learning is possible and has the potential of being a valuable aid in the evaluation of osteoarthritis severity in clinical and research fields as a consistent tool.

CONCLUSION

This study demonstrated that deep learning-based models can effectively perform automated Kellgren-Lawrence grading of knee osteoarthritis from plain radiographs. Using ordinal-aware evaluation metrics, consistent classification performance was achieved across multiple pre-trained architectures, with higher accuracy observed for moderate-to-severe disease stages. Among the evaluated models, VGG-16 showed slightly superior and more stable performance.

Most misclassifications occurred between adjacent KL grades, reflecting the ordinal and progressive nature of osteoarthritis and supporting the clinical plausibility of the proposed framework. In addition to conventional performance metrics, decision curve analysis revealed that the proposed model consistently outperformed default decision strategies across clinically meaningful thresholds, supporting its potential use as a decision-support tool for both early and advanced stages of knee osteoarthritis. These findings suggest that automated KL grading systems may serve as useful decision-support tools by reducing observer-dependent variability in radiographic assessment. Further validation on larger and multi-center datasets is required to enhance clinical generalizability.

Acknowledgment

The authors declared that this study has received no financial support.

REFERENCES

1. Karataş T, Yılmaz E, Polat Ü. Osteoartrit yönetimi, yaşam kalitesi ve hemşirenin destekleyici rolü. Med J SDU. 2022;29(2):265-71.

2. Yıldız K, Çelik S, Taşkın E, Boy F, Aygün Ü. Osteoartrit tanılı hastalarda platelet indekslerinin incelenmesi. Van Saglik Bilim Derg. 2024;17(3):131-5.

3. Bilge A, Ulusoy RG, Üstebay S, Öztürk Ö. Osteoartrit. Kafkas J Med Sci. 2018;8(1):133-42.

4. Misir A, Yildiz KI, Kizkapan TB, Incesoy MA. Kellgren-Lawrence grade of osteoarthritis is associated with change in certain morphological parameters. Knee. 2020;27(3):633-41.

5. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop Relat Res. 2016;474(8):1886-93.

6. Zhao H, Ou L, Zhang Z, Zhang L, Liu K, Kuang J. The value of deep learning-based X-ray techniques in detecting and classifying Kellgren-Lawrence grades of knee osteoarthritis: a systematic review and meta-analysis. Eur Radiol. 2025;35:327-40.

7. Köse Ö, Acar B, Çay F, Yilmaz B, Güler F, Yüksel HY. Inter- and intraobserver reliabilities of four different radiographic grading scales of osteoarthritis of the knee joint. J Knee Surg. 2018;31(3):247-53.

8. Li W, Xiao Z, Liu J, Feng J, Zhu D, Liao J, et al. Deep learning-assisted knee osteoarthritis automatic grading on plain radiographs: the value of multiview X-ray images and prior knowledge. Quant Imaging Med Surg. 2023;13(6):3587-601.

9. Abdullah SS, Rajasekaran MP. Automatic detection and classification of knee osteoarthritis using deep learning approach. Radiol Med. 2022;127:398-406.

10. Olsson S, Akbarian E, Lind A, Razavian A.S, Gordon M. Automating classification of osteoarthritis according to Kellgren-Lawrence in the knee using deep learning in an unfiltered adult population. BMC Musculoskelet Disord. 2021;22(1):844.

11. Kalpana V, Kumar GH. Evaluating the efficacy of deep learning models for knee osteoarthritis prediction based on Kellgren-Lawrence grading system. e-Prime Adv Electr Eng Electron Energy. 2023;5:100266.

12. Kishore VV, Batthala S, Chamarthi JV, Achyutasai C. Knee osteoarthritis prediction driven by deep learning and the Kellgren-Lawrence grading. Proc Eng Sci. 2023;5(3):475-84.

13. Nasef D, Nasef D, Sawiris V, Girgis P, Toma M. Deep learning for automated Kellgren-Lawrence grading in knee osteoarthritis severity assessment. Surgeries. 2024;6(1):3.

14. Vaattovaara E, Panfilov E, Tiulpin A, Niinimäki T, Niinimäki J, Saarakkala S, et al. Kellgren-Lawrence grading of knee osteoarthritis using deep learning: diagnostic performance with external dataset and comparison with four readers. Osteoarthritis Cartilage Open. 2025;7(2):100580.

15. Kılıç Ş. Densenet201+ with multi-scale attention and deep feature engineering for automated Kellgren-Lawrence grading of knee osteoarthritis. Peer J Comput Sci. 2025;11:e3329.

16. Solak FZ. Classification of knee osteoarthritis severity by transfer learning from X-ray images. Karaelmas Fen ve Muhendislik Derg. 2024;14(2):119-33.

17. Mohammed AS, Hasanaath AA, Latif G, Bashar A. Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed X-ray images. Diagnostics. 2023;13(8):1380.

18. Kaggle (Internet). San Francisco (CA): Kaggle Inc.; c2010– (cited 2026 Jan 4). Osteoarthritis Initiative (OAI) full dataset. Available from: https://www.kaggle.com/datasets/tunnguyntrung/oai-full.

19. Yang J, Ji Q, Ni M, Zhang G, Wang Y. Automatic assessment of knee osteoarthritis severity in portable devices based on deep learning. J Orthop Surg Res. 2022;17:540.

20. Dalia Y, Bharath A, Mayya V, Kamath SS. DeepOA: clinical decision support system for early detection and severity grading of knee osteoarthritis. In: Proc IEEE 5th Int Conf Computer, Communication and Signal Processing (ICCCSP); 2021 May 24-25; Chennai, India. p. 250-5.

21. Wahyuningrum RT, Yasid A, Verkerke GJ. Deep neural networks for automatic classification of knee osteoarthritis severity based on X-ray images. In: Proc 8th Int Conf Information Technology (ICIT 2020); 2020 Dec 25-27; Xi'an, China. p. 110-4.

22. Yong CW, Teo K, Murphy BP, Hum YC, Tee YK, Xia K, et al. Knee osteoarthritis severity classification with ordinal regression module. Multimed Tools Appl. 2021;81:41497-509.

23. Ruikar D, Kamble P, Ruikar A, Houde K, Hegadi R. DNN-based knee osteoarthritis severity prediction system: pathologically robust feature engineering approach. SN Comput Sci. 2022;4:58.

24. Jain RK, Sharma PK, Gaj S, Sur A, Ghosh P. Knee osteoarthritis severity prediction using an attentive multi-scale deep convolutional neural network. arXiv. 2021;arXiv:2106.14292.

25. Yunus U, Amin J, Sharif M, Yasmin M, Kadry S, Krishnamoorthy S. Recognition of knee osteoarthritis using YOLOv2 and convolutional neural network-based classification. Life. 2022;12:1126.

26. Raisuddin AM, Nguyen HH, Tiulpin A. Deep semi-supervised active learning for knee osteoarthritis severity grading. In: Proc IEEE Int Symp Biomedical Imaging (ISBI); 2022 Mar 28-31; Kolkata, India. p. 1-5.

Bozok Tıp Derg 2026;16(1):115-125
Bozok Med J 2026;16(1):115-125

27. Wang Y, Bi Z, Xie Y, Wu T, Zeng X, Chen S, et al. Learning from highly confident samples for automatic knee osteoarthritis severity assessment: data from the Osteoarthritis Initiative. IEEE J Biomed Health Inform. 2022;26:1239-50.

28. Alshamrani HA, Rashid M, Alshamrani SS, Alshehri AHD. OsteoNeT: an automated system for predicting knee osteoarthritis from X-ray images using transfer-learning-based neural networks. Healthcare. 2023;11:1206.