EJMI

DOI: 10.14744/ejmi.263080
EJMI 2026;10(1):80–83

Research Article

Interobserver Reliability of the Kellgren–Lawrence Classification in Knee Osteoarthritis: A Comparison Between Orthopedic Surgeons and Artificial Intelligence

Şafak Sayar[1], Mustafa Boz[1], Yasemin Begüm Topkarcı[2], Suat Batar[1], Necdet Demir[1]

[1]Department of Orthopaedics and Traumatology, Biruni University Faculty of Medicine Hospital, Istanbul, Türkiye
[2]Faculty of Medicine, Gazi University, Ankara, Türkiye

Abstract

Objectives: To evaluate the interobserver reliability of the Kellgren–Lawrence (KL) classification among orthopedic surgeons and to compare their assessments with artificial intelligence (AI) systems.

Methods: One hundred anteroposterior weight-bearing knee radiographs from patients aged 65 years and older were retrospectively analyzed. Four orthopedic surgeons and two AI systems independently graded all radiographs according to the KL classification and were blinded to clinical information and to each other's evaluations. Interobserver agreement was assessed using quadratically weighted Cohen's kappa and intraclass correlation coefficients (ICC).

Results: Interobserver agreement among orthopedic surgeons demonstrated good reliability (mean weightedκ=0.780; ICC=0.784). Agreement between the orthopedic consensus and ChatGPT was moderate (κ=0.481), whereas Gemini demonstrated moderate-to-good agreement (κ=0.561). Agreement between the two AI systems was also moderate (κ=0.484).

Conclusion: The KL classification demonstrated good reliability among orthopedic surgeons. AI systems demonstrated moderate agreement with orthopedic experts and may serve as supportive screening tools rather than as diagnostic replacements.

Keywords: Artificial Intelligence, Interobserver Reliability, Kellgren–Lawrence Classification, Knee Osteoarthritis, Radiographic Grading

Cite This Article: *Sayar Ş, Boz M, Topkarcı YB, Batar S, Demir N. Interobserver Reliability of the Kellgren–Lawrence Classification in Knee Osteoarthritis: A Comparison Between Orthopedic Surgeons and Artificial Intelligence. EJMI 2026;10(1):80–83.*

Knee osteoarthritis is among the most common degenerative joint diseases worldwide and a leading cause of disability, particularly among older adults.[1] Radiographic evaluation continues to be an important part of assessing disease severity.

The Kellgren–Lawrence(KL)classification, first described in 1957, is the most widely used radiographic grading system for knee osteoarthritis.[2] Although it is widely used, concerns remain regarding its observer dependency and variability among clinicians.[3,4]

Previous studies have reported moderate to good interobserver agreement for the KL grading scale.[4,5] With the increasing integration of artificial intelligence (AI) into medi-

Address for correspondence: Şafak Sayar, MD. Department of Orthopaedics and Traumatology, Biruni University Faculty of Medicine Hospital, Istanbul, Türkiye
Phone: +90 533 439 10 01 **E-mail:** safaksayar@gmail.com

Submitted Date: February 28, 2026 **Accepted Date:** March 19, 2026

cal imaging, automated grading systems have emerged.[6,7] Nevertheless, insufficient research has been conducted on the compatibility between AI systems and skilled orthopedic surgeons.

The aim of this study was to compare orthopedic surgeons' evaluations with those of AI systems and to assess the interobserver reliability of the KL classification.

Methods

Ethics committee approval was obtained from Biruni University Ethics Committee (approval no: 2024-BİAEK/18-65, date: 23.02.2026). The study was conducted in accordance with the principles of the Declaration of Helsinki. This retrospective study assessed the reliability and agreement of a widely used radiographic classification system for knee osteoarthritis. Our study included 100 anonymized anteroposterior weight-bearing knee radiographs from patients aged 65 years and older.

To maintain homogeneity within the study population and focus on primary degenerative osteoarthritis, only patients aged 65 years and older were included. Previous knee surgery, inflammatory arthropathies, post-traumatic osteoarthritis, congenital deformities, and poor image quality were among the exclusion criteria. Only one knee per patient was included.

Ethics Statement

This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained from the Biruni University Ethics Committee prior to the initiation of the study. Because the study was retrospective and used fully anonymized radiographic data, the ethics committee waived the requirement for informed consent.

All radiographs were de-identified before analysis, and no patient-identifiable information was accessible to any of the observers, including the AI systems.

Radiographic Evaluation

Radiographs were independently evaluated by four orthopedic surgeons and two AI systems (ChatGPT 5.2 and Gemini 3). To reduce recall bias, the order of radiographs was randomly assigned. The orthopedic surgeons had experience in musculoskeletal imaging and knee osteoarthritis management. Images were uploaded directly, in high-resolution JPEG format, to the AI interfaces, which evaluated the radiographs based solely on predefined classification criteria without access to any clinical or demographic information. To ensure standardization, both AI systems received the following prompt: "Evaluate the following knee radiograph according to the Kellgren-Lawrence classification (Grades 0-4).

All observers were blinded to clinical data and to each other's evaluations. Radiographs were categorized using grades 0–4 in accordance with the KL classification system. Prior to evaluation, each observer received a standardized written description of the KL grading criteria. No clinical data, patient history, physical examination findings, or previous imaging results were available to any observer. A consensus among surgeons was determined using majority grading.

Statistical Analysis

Statistical analyses were performed using Python (Python Software Foundation, Wilmington, Delaware, USA). Interobserver agreement was evaluated using the weighted Cohen's kappa coefficient. The primary outcome of the study was the interobserver reliability of the KL classification, both among orthopedic surgeons and between orthopedic surgeons and AI systems.

Since the KL classification represents an ordinal categorical scale, weighted Cohen's kappa (κ) statistics were used as the primary measure of interobserver agreement. Pairwise weighted kappa values were calculated for all pairs of observers. To provide an overall measure of agreement among the orthopedic surgeons, the mean weighted kappa value was reported. Agreement between the orthopedic surgeons and each AI system was assessed by comparing each AI system's classifications with the surgeons' consensus grading.

Kappa values were interpreted according to the criteria proposed by Landis and Koch: values <0.20 were considered poor agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, and >0.80 very good agreement.[8]

As a secondary analysis, intraclass correlation coefficients (ICCs) were calculated using a two-way random-effects model and the absolute-agreement definition to further assess the reliability of the KL grading among all observers. Results were reported with 95% confidence intervals (CI).

All statistical analyses were performed using appropriate statistical software. A p-value of <0.05 was considered statistically significant.

Results

Interobserver agreement among the four orthopedic surgeons ranged between κ=0.712 and κ=0.867, with a mean weighted kappa of 0.780. This corresponds to good interobserver reliability according to the Landis and Koch criteria (Table 1).

In addition, the intraclass correlation coefficient (ICC, two-way random, absolute agreement) calculated for the four surgeons was 0.784. This further confirms a good overall level of agreement among orthopedic surgeons.

Agreement between the orthopedic consensus and ChatGPT 5.2 was moderate (κ=0.481), whereas Gemini 3 showed moderate-to-good agreement (κ=0.561) (Fig. 1).

Agreement between AI systems (ChatGPT 5.2 and Gemini 3) was moderate (κ=0.484) (Table 2).

Table 1. Interobserver agreement among orthopedic surgeons

Observer pair	Weighted Kappa
Surgeon1–Surgeon2	0.745
Surgeon1–Surgeon3	0.712
Surgeon1–Surgeon4	0.867
Surgeon2–Surgeon3	0.713
Surgeon2–Surgeon4	0.838
Surgeon3–Surgeon4	0.804

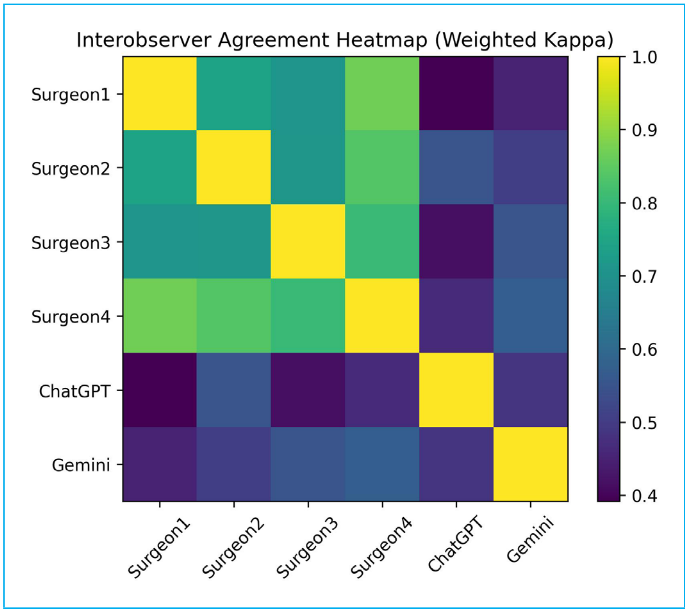

Figure 1. Interobserver agreement heatmap based on weighted kappa values. Heatmap demonstrating pairwise interobserver reliability for Kellgren–Lawrence grading of knee osteoarthritis among four orthopedic surgeons and two artificial intelligence systems. Color intensity represents the magnitude of agreement (weighted κ), ranging from low agreement (dark colors) to perfect agreement (bright yellow). Diagonal values indicate self-agreement (κ=1.0), while off-diagonal values reflect agreement between different observers. 137x121 mm (300x300 DPI).

Table 2. Agreement between orthopedic consensus and AI systems

Comparison	Weighted Kappa
Consensus–ChatGPT	0.481
Consensus–Gemini	0.561

Discussion

The main finding of this study is that the KL classification has good interobserver reliability among experienced orthopedic surgeons, whereas current AI systems only achieve moderate agreement with human experts. The observed mean weighted kappa of 0.780 and ICC of 0.784 are consistent with previously reported reliability ranges in the literature.[4,5]

The relatively high agreement among surgeons in this cohort may be explained by the inclusion of patients aged 65 years and older, in whom radiographic degenerative changes are typically more pronounced. Clear osteophyte formation and joint space narrowing may reduce grading uncertainty compared with early-stage disease.[3,9]

AI systems showed moderate agreement with the orthopedic consensus. Although AI performance did not reach surgeon-level reliability, the findings suggest that it could be useful as a support tool for radiographic standardization and large-scale screening. Similar AI-based approaches have demonstrated promising results in automated radiography grading,[6,7] though performance remains dependent on training datasets and algorithmic design.

It is important to note that the KL classification is partially observer-dependent, as it is based on visual interpretation of osteophyte formation and joint space narrowing.[3] As a result, variability between human observers and AI systems may reflect intrinsic limitations of the classification system rather than solely technological shortcomings.

Radiographic grading alone should not be used to make clinical decisions without consideration of patient symptoms, functional status, and clinical examination findings.[10]

Limitations

This study is limited by its retrospective design and inclusion of patients aged 65 years or older, which may reduce generalizability to younger populations. Clinical correlation and advanced imaging modalities were not evaluated. AI systems were assessed as static observers, and their performance may vary with future algorithmic developments.

Conclusion

The KL classification demonstrates good interobserver reliability among orthopedic surgeons in an elderly population. AI systems show moderate agreement with human observers, indicating their potential as supportive tools in radiographic assessment. However, AI systems have yet to achieve surgeon-level agreement and should supplement, rather than replace, expert clinical evaluation.

Disclosures

Ethics Committee Approval: Ethics committee approval was obtained from Biruni University Ethics Committee (approval no: 2024-BİAEK/18-65, date: 23.02.2026).

Informed Consent: Written informed consent was obtained from all patients included in the study.

Authors' contributions: Concept – SS, MB; Design – SS, MB; Supervision – SS; Materials – SS, SB, ND; Data Collection and/or Processing – MB, YBT, SB; Analysis and/or Interpretation – SS, YBT; Literature Review – YBT, ND; Writing – SS; Critical Review – MB, SB, ND.

Conflict of Interest: The authors declare that there is no conflict of interest.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Use of AI for Writing Assistance: Artificial intelligence tools were used for statistical analysis. The authors reviewed, verified, and approved all results. The authors take full responsibility for the accuracy and integrity of the data and analyses presented in the manuscript.

Peer-review: Externally peer-reviewed.

References

1. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet 2019;393(10182):1745–1759. [CrossRef]
2. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16(4):494–502. [CrossRef]
3. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren–Lawrence classification of osteoarthritis. Clin Orthop Relat Res 2016;474(8):1886–1893. [CrossRef]
4. Damen J, Schiphof D, Wolde ST, Cats HA, Bierma-Zeinstra SM, Oei EH. Inter-observer reliability for radiographic assessment of early osteoarthritis features: the CHECK (cohort hip and cohort knee) study. Osteoarthritis Cartilage 2014;22(7):969–74. [CrossRef]
5. Wright RW; MARS Group. Osteoarthritis Classification Scales: Interobserver Reliability and Arthroscopic Correlation. J Bone Joint Surg Am 2014;96(14):1145–51. [CrossRef]
6. Antony J, McGuiness K, O' Connor N, Moran K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. Med Image Anal 2017;39(1):1–13.
7. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Sci Rep 2018;8(1):1727. [CrossRef]
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174. [CrossRef]
9. Guermazi A, Hayashi D, Eckstein F, Hunter DJ, Duryea J, Roemer FW. Imaging of osteoarthritis. Rheum Dis Clin North Am 2013;39(1):67–105. [CrossRef]
10. Altman RD, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, et al. The American College of Rheumatology criteria for osteoarthritis of the knee. Arthritis Rheum 1986;29(8):1039–1049. [CrossRef]