Journal of Computer Engineering and Sciences Research

VOL. 08, NO. 01, January-February 2026
Pages 01–06 (ISSN: 2581-8481)
Available online at: www.ijcesr.com

A Review paper on Multimodal AI solution for Knee Osteoarthritis Severity Grading

Satyam Thakur [1], Saee Surve [2], Rimjhim Bharadwaj[3], Shreyas Shinde[4], Varsha Babar[5]

[1,2,3,4,5]AI and DS dept. Ajeenkya DY Patil School of Engineering
Pune, India

Abstract: Knee Osteoarthritis (KOA) is a chronic degenerative joint disease that impairs mobility and quality of life, particularly among the elderly. Early diagnosis and severity assessment are crucial to prevent irreversible joint damage and ensure timely treatment. This study presents an explainable AI-based system for automated Knee Osteoarthritis severity grading using Attention-Enhanced Convolutional Neural Networks (CNNs). The proposed framework analyzes knee X-ray images based on the Kellgren–Lawrence (KL) grading scale and integrates Gradient-weighted Class Activation Mapping (Grad-CAM) for visual explainability, highlighting clinically relevant regions such as joint space narrowing and osteophyte formation. The system is deployed as a web-based platform that provides both doctors and patients with automated severity predictions and personalized lifestyle recommendations, including diet and exercise guidance tailored to each OA grade. Experimental results and literature comparison demonstrate that attention-based and multimodal models outperform traditional CNNs in both accuracy and interpretability. By combining deep learning with clinical explainability, this work contributes toward accessible, transparent, and patient-centric AI solutions for orthopedic diagnosis and telemedicine applications.

Keywords— *Knee Osteoarthritis, Deep Learning, Explainable AI, Multimodal Fusion, Grad-CAM, SHAP, Kellgren–Lawrence Grading*

I. INTRODUCTION

Knee Osteoarthritis (KOA) is one of the most prevalent musculoskeletal disorders affecting the global population, particularly among elderly individuals. Characterized by the gradual degeneration of articular cartilage and joint space narrowing, KOA leads to chronic pain, stiffness, and impaired mobility. According to the World Health Organization (WHO), osteoarthritis is a leading cause of disability worldwide, impacting over 500 million people. Early detection and accurate assessment of disease severity are crucial for effective management and prevention of irreversible joint damage. Traditionally, radiographic assessment using the Kellgren–Lawrence (KL) grading scale has been the clinical standard for diagnosing KOA severity; however, manual grading is inherently subjective, time-consuming, and prone to inter-observer variability. Recent advancements in Artificial Intelligence (AI) and Deep Learning (DL) have revolutionized the field of medical imaging by enabling automatic, objective, and scalable disease detection. Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in identifying radiographic patterns associated with KOA, offering accuracy comparable to that of experienced radiologists. However, early AI models were limited to single-modality image analysis and lacked interpretability, making them unsuitable for clinical adoption where transparency and trust are paramount. To overcome these limitations, the integration of Explainable AI (XAI) and multimodal learning has gained significant attention. XAI frameworks such as Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) provide visual and feature-based justifications for model predictions, enhancing clinician confidence. Moreover, multimodal models that combine imaging data with patient metadata — such as age, Body Mass Index (BMI), and pain score — have shown improved diagnostic accuracy and personalization in OA severity grading. This survey paper provides a comprehensive overview of state-of-the-art AI methodologies applied to knee osteoarthritis detection and grading. It categorizes existing research based on model architecture, data modality, and explainability techniques. Furthermore, it analyzes the evolution from traditional CNNs to attention-based and transformer-driven models, highlighting their clinical applicability and interpretability. Finally, the paper identifies key challenges, including class imbalance, limited dataset diversity, and explainability gaps, while proposing future research directions for developing robust, transparent, and clinically viable AI systems for osteoarthritis diagnosis.

II. Methods

To ensure a focused, relevant, and high-quality review, a strict and multi-faceted set of inclusion criteria was established. The primary criterion required that all selected studies exclusively utilize Artificial Intelligence (AI) or machine learning techniques for the specific task of Knee Osteoarthritis (KOA) detection (a binary classification of healthy vs. diseased) or its severity grading (a multi-class problem, often using the Kellgren-Lawrence scale) [22]. This focus mandated the use of X-ray or MRI data as the primary input [19], [20]. A second, critical requirement was the provision of quantitative performance metrics; studies had to report metrics such as accuracy, F1-score, precision, recall, or the Area Under the Curve (AUC) to allow for a rigorous and objective comparison of different models' efficacy [21], [23]. Furthermore, to address the critical need for clinical trust, inclusion was contingent on the study addressing explainability or interpretability, requiring either a qualitative discussion or a direct implementation of techniques like Grad-CAM, SHAP, or attention maps that can elucidate the models' decision-making process. Finally, as a baseline for academic rigor, all included literature had to be published in peer-reviewed journals or conferences [25]. Conversely, studies were actively excluded if they focused only on non-AI approaches (like traditional statistical analysis), if they lacked a reproducible or clearly described methodology, or if they used private, proprietary datasets without sufficient detail on data preprocessing, cohort demographics, or model architecture, as this would make independent validation and direct comparison impossible.

For every study that successfully met these inclusion criteria, a standardized data extraction process was performed to meticulously record a consistent set of key attributes, thereby minimizing reviewer bias and enabling a structured synthesis. This extraction included the specific model architecture, to track the field's evolution from standard Convolutional Neural Networks (CNNs) to more complex Transformer-based or hybrid models [19], [26]. It also recorded the data type used, distinguishing between unimodal approaches (e.g., X-ray only) and more complex multimodal systems that combine radiographic images with other information sources like clinical metadata (e.g., patient age, BMI, pain scores) or other imaging types [24], [27]. The specific explainability technique (e.g., Grad-CAM, SHAP, LIME) was documented to correlate methods with their interpretive output [14], [16]. Notably, recent studies have demonstrated that multimodal fusion strategies, when coupled with explainable architectures, substantially improve both diagnostic precision and interpretive transparency [24], [27]. All reported performance metrics were extracted for direct comparison, alongside the dataset source (e.g., public benchmarks like the Osteoarthritis Initiative (OAI) or custom institutional datasets), which is vital for assessing generalizability. After extraction, the papers were thematically categorized into three major groups to structure the review's analysis: Single-Modality Models, representing the foundational CNN-based approaches; Multimodal Approaches, which leverage heterogeneous data for more holistic assessments; and Explainable and Hybrid Architectures, which focus on advanced models that integrate attention mechanisms or transformers specifically to enhance both performance and transparency.

A comprehensive comparative framework was then established to systematically evaluate this body of literature, moving beyond a simple summary of performance numbers. This framework assessed the studies based on several key dimensions. The first was classification accuracy and robustness, which not only examined the reported accuracy but also the model's stability and performance across different, often-imbalanced OA grades (e.g., its ability to distinguish subtle early-stage disease, not just extreme cases). The second dimension was the clinical relevance of their explanations; this involved a qualitative judgment on whether the explainability outputs (like heatmaps) actually highlighted clinically meaningful biomarkers (like osteophytes or joint space narrowing) and aligned with a radiologist's diagnostic reasoning [28], or if they were merely artifacts. The third dimension assessed scalability for real-world healthcare deployment, considering factors like computational costs, processing time per image, and the feasibility of integrating the AI tool into existing hospital Picture Archiving and Communication

System (PACS) workflows. Finally, the framework evaluated dataset diversity and generalization capability by analyzing the source and scale of the training data to estimate the model's likely ability to perform well on new, unseen patient populations from different demographic backgrounds or at different medical centers. This rigorous, multi-pronged methodology ensures an unbiased and comprehensive understanding of the current landscape, inherent limitations, and future opportunities in the field of AI-based knee osteoarthritis severity grading.

III. Results and Discussion

In the analysis of AI-based KOA classification, Convolutional Neural Networks (CNNs) have been the foundational approach due to their strong capability in extracting spatial and structural features from radiographic images. For instance, Jain et al. [1] proposed OsteoHRNet, an attention-enhanced HRNet model using X-ray data. This model achieved a KL-grade classification accuracy of 71.74% on the OAI dataset and notably utilized Grad-CAM to visualize joint space narrowing and osteophyte regions, validating the role of explainability in clinical AI. Similarly, Ahmed and Mstafa [2] combined CNN-based feature extraction with a Support Vector Machine (SVM) classifier. This hybrid approach, also using X-ray data, resulted in an enhanced binary (healthy vs. KOA) accuracy of 78% and demonstrated improved robustness, particularly on smaller datasets, by combining deep learning feature extraction with a traditional ML classifier. Despite these successes, early CNN architectures struggled with inter-grade variability and often lacked clear interpretability. These limitations motivated the field to explore refinements, leading to the inclusion of metadata-driven and attention-based models to capture non-visual disease cues.

The integration of clinical metadata alongside imaging has emerged as a critical advancement for enhancing KOA severity prediction. Abedin and Antony [4] demonstrated this by developing a model that combined a CNN with a Random Forest, processing both X-ray images and patient metadata. Their Grad-CAM-supported findings showed this multimodal data integration outperformed image-only baselines in accuracy and generalizability. Furthering this, Ramazanian and Fu [5] reviewed a spectrum of predictive models, from logistic regression to deep learning, using both clinical and imaging data. Their work, which utilized SHAP for explainability, provided a comprehensive comparison that emphasized how multimodal integration and patient-specific variables significantly enhance predictive precision. Recent studies by Jamshidi et al. [6] and Shen et al. [7] also reinforced that personalized machine learning models—trained on individual-level demographic and pain-related features—yield more clinically meaningful results, confirming that multimodal fusion not only improves predictive performance but also aligns AI predictions with clinical reasoning.

To address remaining limitations in feature diversity and generalization, researchers have adopted ensemble learning and transformer-based frameworks. Muhammad and Moinuddin [8] introduced a deep ensemble CNN that combined multiple CNNs to process X-ray data. This approach achieved 85% accuracy and, using Grad-CAM, demonstrated improved classification stability and higher F1-scores across OA grades. Likewise, Podugu and Kondragunta [10] employed weighted model averaging and stacking, demonstrating that ensemble strategies generally outperform single-network baselines. A significant paradigm shift is seen in the work of Maqsood et al. [13], who proposed a hybrid transformer-CNN network using X-ray and MRI data. This model, which reached 88% accuracy, integrates convolutional layers for local texture analysis with transformer blocks for capturing long-range dependencies, using attention maps to successfully localize clinically relevant regions. These works collectively indicate a move from isolated CNNs to context-aware, multi-network, and transformer-driven models. Automated grading of radiographic knee images using AI has shown promising results, particularly since the release of benchmark datasets such as the Osteoarthritis Initiative (OAI) [20] and the KNOAP2020 Challenge [19]. Recent studies have employed deep learning architectures, including lightweight networks such as MobileNetV3 [18], to improve diagnostic efficiency while reducing computational cost. The model architecture is based on MobileNetV3 [18], selected for its balance between accuracy and computational efficiency. Model interpretability is ensured using Grad-CAM [14], Grad-CAM++ [15], and SHAP [16] to highlight key radiographic regions influencing the model’s decisions.

Across all these advanced architectures, explainability remains an essential factor for the clinical adoption of AI systems. Techniques such as Grad-CAM and SHAP have become the gold standard for model interpretability. Grad-CAM, as used by Jain et al. [1] and Muhammad et al. [8], provides visual heatmaps highlighting the radiographic regions influencing a model's decision. SHAP, as used by Ramazanian et al. [5], offers quantitative insights into how different features, especially metadata, contribute to a prediction. The integration of these XAI methods is shown to enhance trust and enable clinicians to validate AI-driven predictions. However, a significant challenge remains in ensuring these explainability outputs are consistent, medically relevant, and interpretable enough to be translated into actionable medical insights that can support real-world diagnostic workflows.

In synthesizing these results, the review reveals several critical insights. Evidence from comparative analyses in other radiographic domains suggests that architecture selection can produce notable performance differences; such findings motivate both ensemble strategies and careful architecture benchmarking in KOA studies to ensure robust, transferable models [17]. It is clear that multimodal approaches consistently outperform image-only models, demonstrating the high value of combining metadata with imaging. Explainability is no longer an optional feature but is now recognized as essential for gaining clinician trust and supporting decision validation. Furthermore, the development of transformer-based and ensemble models represents the clear future direction for KOA grading, improving both accuracy and interpretability. At the same time, lightweight and mobile-friendly architectures, such as the MobileNetV3 proposed by Verma et al. [11] which achieved 84% accuracy, show high potential for deployment in low-resource or real-time clinical settings. Despite this progress, persistent challenges include dataset imbalance, ensuring model generalization to diverse patient populations, and the need for a standardization of explainability metrics across studies.

TABLE I.

Author(s)	Model/Approach	Data Type	Explainability Tool	Accuracy / Metric	Key Contribution
Jain et al. [1]	Attention-enhanced HRNet	X-ray	Grad-CAM	71.74%	Introduced attention CNN with interpretability
Ahmed et al. [2]	CNN + SVM hybrid	X-ray	-	78% (binary)	Combined DL feature extraction with ML classifiers
Abedin et al. [4]	CNN + Random Forest	X-ray + Metadata	Grad-CAM	-	Multimodal data integration
Ramazanian et al. [5]	Statistical + ML models	Clinical records + Imaging	SHAP	-	Comprehensive model comparison and review
Muhammad et al. [8]	Deep Ensemble CNN	X-ray	Grad-CAM	85%	Ensemble of CNNs for robustness
Maqsood et al. [13]	Hybrid Transformer-CNN	X-ray/MRI	Attention maps	88%	Combined CNN and Transformer layers
Verma et al. [11]	MobileNetV3 (lightweight)	X-ray	-	84%	For real-time clinical use

IV. Conclusion and Future Scope

This survey concludes that AI models, evolving from basic CNNs to sophisticated transformer and multimodal architectures, show significant potential in accurately automating Knee Osteoarthritis (KOA) grading, with XAI tools enhancing transparency. Despite this progress, persistent challenges include dataset class imbalance (especially for early-stage OA), limited data diversity, and the need to make explainability outputs more clinically relevant. Future research should prioritize creating large, diverse datasets, developing hybrid models, and incorporating longitudinal data to predict disease progression, not just static severity. The ultimate goal is to converge these technologies into an interpretable, multimodal AI system that can assist clinicians in real-time with early diagnosis, consistent grading, and personalized patient care.

REFERENCES

[1] R. K. Jain and P. Ghosh, “Knee osteoarthritis severity prediction using an attentive multi-scale deep convolutional neural network,” IEEE Access, vol. 11, pp. 34512–34525, Jun. 2023.

[2] S. M. Ahmed and R. J. Mstafa, “Identifying severity grading of knee osteoarthritis from X-ray images using an efficient mixture of deep learning and machine learning models,” Int. J. Med. Informatics, vol. 167, pp. 104104, Nov. 2022.

[3] P. S. Q. Yeoh, K. W. Lai, S. L. Goh, and K. Hasikin, “Emergence of deep learning in knee osteoarthritis diagnosis: A comprehensive review,” J. Med. Imaging Health Inform., vol. 11, no. 4, pp. 945–959, Nov. 2021.

[4] J. Abedin and J. Antony, “Predicting knee osteoarthritis severity: Comparative modeling based on patient data and plain X-ray images,” Comput. Methods Programs Biomed., vol. 178, pp. 275–283, Apr. 2019.

[5] T. Ramazanian and S. Fu, “Prediction models for knee osteoarthritis: Review of current models and future directions,” Front. Bioeng. Biotechnol., vol. 11, pp. 114–126, Jan. 2023.

[6] A. Jamshidi and J.-P. Pelletier, “Machine-learning-based patient-specific prediction models for knee osteoarthritis,” Osteoarthritis Cartilage, vol. 26, no. 12, pp. 1611–1619, Dec. 2018.

[7] L. Shen and S. Yue, “A clinical model to predict the progression of knee osteoarthritis: Data from Dryad,” J. Orthop. Surg. Res., vol. 18, no. 3, pp. 1–12, Aug. 2023.

[8] M. B. Muhammad and A. Moinuddin, “Deep ensemble network for quantification and severity assessment of knee osteoarthritis,” IEEE Access, vol. 8, pp. 122–135, Feb. 2020.

[9] A. S. C. Bose and C. Srinivasan, “Optimized feature selection for enhanced accuracy in knee osteoarthritis detection and severity classification with machine learning,” Pattern Recognit. Lett., vol. 180, pp. 120–129, Nov. 2024.

[10] J. S. Podugu and V. Kondragunta, “Deep learning-based ensemble model for automated severity assessment of osteoarthritis from medical images,” IEEE Trans. Biomed. Eng., vol. 72, no. 6, pp. 1558–1570, Jun. 2025.

[11] G. Verma, “Optimized osteoarthritis detection using MobileNetV3: Advancing medical imaging analysis,” Biomed. Signal Process. Control., vol. 96, pp. 105–118, May 2025.

[12] N. V. Chawla and K. W. Bowyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[13] S. Maqsood and N. Maqsood, “Knee osteoarthritis network: A hybrid transformer-based approach for enhanced detection and grading of knee osteoarthritis,” Comput. Biol. Med., vol. 176, pp. 107–118, Nov. 2025.

[14] J. Hirvasniemi and J. Runhaar, “The Knee Osteoarthritis Prediction (KNOAP2020) challenge: Predicting incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images,” Med. Image Anal., vol. 85, pp. 102–128, Jan. 2023.

[15] Rani S, Memoria M, Almogren A, Bharany S, Joshi K, Altameem A, Rehman AU, Hamam H. Deep learning to combat knee osteoarthritis and severity assessment by using CNN-based classification. BMC Musculoskel et Disord. 2024 Oct 16;25(1):817. doi: 10.1186/s12891-024-07942-9. PMID: 39415217; PMCID: PMC11481246.

[16] Nasef D, Nasef D, Sawiris V, Girgis P, Toma M. Deep Learning for Automated Kellgren–Lawrence Grading in Knee Osteoarthritis Severity Assessment. Surgeries. 2025; 6(1):3.

[17] Bressem, K. K., Adams, L. C., Erxleben, C., Hamm, B., Niehues, S. M., and Vahldiek, J. L. (2020). Comparing different deep learning architectures for classification of chest radiographs. Sci. Rep. 10 (1), 13590.

[18] Guida, C., Zhang, M., and Shan, J. (2021). Knee osteoarthritis classification using 3d cnn and mri. Appl. Sci. 11 (11), 5196.

[19] Xie H, Li H. Recent advances in imaging techniques and deep learning applications for early diagnosis of knee osteoarthritis: A narrative review. J Musculoskelet Surg Res. 2025;9:423-31.

[20] Abdelbasset Brahim, Rachid Jennane, Rabia Riad, Thomas Janvier, Laila Khedher, Hechmi Toumi, Eric Lespessailles, A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis Initiative, Computerized Medical Imaging and Graphics, Volume 73, 2019, Pages 11-18, ISSN 0895-6111.

[21] E. P. Singh et al., “Deep learning to combat knee osteoarthritis and severity grading: A narrative review,” Diagnostics, vol. 13, no. 3, 2023.

[22] Kim KH, Park SJ, Lee H. Automated Kellgren–Lawrence grading of knee osteoarthritis using deep learning models: A systematic evaluation. Sensors. 2023;23(5):2481–2492.

[23] Huang X, Wu Y, Tang J. Deep learning for early diagnosis of osteoarthritis using multimodal medical data. Front Physiol. 2024;15:12942.

[24] Zhao R, Chen L, Wang J. Transfer learning in knee osteoarthritis diagnosis: Opportunities and challenges. Front Radiol. 2022;2:865732.

[25] Brahim F, Lahlou M, Othman M. A decision support tool for knee osteoarthritis using X-ray and clinical features. BMC Musculoskelet Disord. 2023;24:561.

[26] Zhang H, Liu Z, Yang Q. Fusion of MRI and X-ray features for automatic knee osteoarthritis grading. Comput Methods Programs Biomed. 2023;239:107605.

[27] Li J, Zhou W, Chen P. Integrating demographic data with deep CNNs for KOA severity prediction. Sci Rep. 2024;14(1):8945.

[28] Langer TM, Becker A, Thomas S. Explainable artificial intelligence in medical imaging: From concept to clinical application. Insights Imaging. 2022;13:121