INSIGHTS-JOURNAL OF
LIFE AND SOCIAL SCIENCES

INSIGHTS-JLSS
Insights-Journal of Life and
Social Sciences

ARTIFICIAL INTELLIGENCE IN MUSCULOSKELETAL RADIOLOGY: A SYSTEMATIC REVIEW OF DIAGNOSTIC ACCURACY AND CLINICAL INTEGRATION

Original Article

Haiya Mahmood[1]*, Sidra Haq[2], Komal Abrar[3], Linta Naveed[4], Anusha Mandhan[5], Zunaira Rizwan[6]

[1]Student, Foundation University Medical College, Pakistan.
[2]Lecturer, King Edward Medical University, Lahore, Pakistan.
[3]Senior Lecturer (MIT), Pakistan.
[4]2nd Year MBBS Student, Jinnah Medical and Dental College, Karachi, Pakistan.
[5]Medical Officer, Chandka Medical College, Larkana, Pakistan
[6]RMO, Tabba Heart Institute, Karachi, Pakistan.

Corresponding Author: Haiya Mahmood, Student, Foundation University Medical College, Pakistan, haiyamahmood@gmail.com

Conflict of Interest: None Grant Support & Financial Support: None

Acknowledgment: The authors would like to acknowledge the contributions of all researchers whose work formed the basis of this systematic review. Special thanks are extended to the institutional library staff for their assistance with database access and to the reviewers who provided critical insights during the study selection and data extraction process.

ABSTRACT

Background: Artificial intelligence (AI) has rapidly emerged as a transformative tool in musculoskeletal radiology, offering the potential to enhance diagnostic accuracy, reduce radiologist workload, and streamline clinical workflows. Despite numerous studies exploring AI applications across various imaging modalities, there remains limited consensus on their diagnostic reliability and integration into routine clinical practice. This gap highlights the need for a comprehensive evaluation of AI's effectiveness in musculoskeletal imaging.

Objective: This systematic review aims to evaluate the diagnostic performance, clinical benefits, and limitations of AI tools in detecting musculoskeletal conditions through radiographic imaging.

Methods: A systematic review was conducted in accordance with PRISMA guidelines. Literature searches were performed in PubMed, Scopus, Web of Science, and the Cochrane Library for studies published between 2018 and 2024. Eligible studies included randomized controlled trials, cohort studies, and cross-sectional designs evaluating AI in musculoskeletal radiology. Inclusion criteria encompassed human studies with comparative diagnostic data. Data were extracted using a standardized form and assessed for bias using the Cochrane Risk of Bias Tool and Newcastle-Ottawa Scale. Due to heterogeneity in study designs and outcomes, a qualitative synthesis was conducted.

Results: Eight studies met inclusion criteria, encompassing a range of musculoskeletal conditions such as fractures, osteoarthritis, and skeletal maturity assessment. AI models, primarily deep learning algorithms, consistently demonstrated high diagnostic performance with sensitivity and specificity exceeding 85% and AUC values often above 0.90. Despite strong accuracy, methodological variability and limited external validation were noted across studies.

Conclusion: AI tools show strong potential in musculoskeletal radiology, demonstrating diagnostic performance comparable to expert radiologists. However, real-world clinical implementation remains limited by variability in study methods and generalizability. Further large-scale, multicenter studies are necessary to confirm clinical utility and integration strategies.

Keywords: Artificial Intelligence, Musculoskeletal Radiology, Diagnostic Accuracy, Deep Learning, Systematic Review, Medical Imaging.

INTRODUCTION

Artificial intelligence (AI) has emerged as a transformative force in radiology, offering the potential to enhance diagnostic accuracy, reduce clinician workload, and improve patient outcomes. Within the domain of musculoskeletal radiology, the application of AI tools—particularly deep learning and machine learning algorithms—has expanded rapidly, with an increasing number of studies exploring their efficacy in interpreting radiographic imaging such as X-rays, MRI, and CT scans. Musculoskeletal disorders, including osteoarthritis, fractures, and soft tissue injuries, are prevalent globally and contribute significantly to disability, pain, and healthcare utilization. For instance, osteoarthritis alone affects over 300 million individuals worldwide, posing substantial clinical and socioeconomic burdens (1,2). Accurate and timely diagnosis is essential for optimal management, yet limitations in radiologist availability and inter-observer variability often hinder effective care. In recent years, numerous AI-based diagnostic tools have demonstrated promising performance in detecting musculoskeletal abnormalities, in some cases rivaling or surpassing human experts (3). However, existing studies vary considerably in methodology, outcome measures, and clinical integration strategies. While some algorithms have shown excellent diagnostic accuracy in controlled research environments, their real-world applicability and generalizability remain unclear (4-6). Moreover, there is limited consensus on the comparative performance of AI systems versus traditional clinical workflows, and questions persist about their regulatory approval, ethical implications, and integration into routine practice (7). Despite the growing body of literature, no comprehensive synthesis has yet thoroughly evaluated both the diagnostic performance and clinical impact of AI in musculoskeletal radiology (8). This gap underscores the need for a systematic review to consolidate current evidence, assess methodological quality, and identify areas for future research.

This review addresses the following research question based on the PICO framework: In patients undergoing radiographic imaging for musculoskeletal conditions (Population), how do AI-based diagnostic tools (Intervention) compare to conventional diagnostic methods or expert radiologist interpretation (Comparison) in terms of diagnostic accuracy and clinical integration outcomes (Outcome)? The primary objective is to evaluate the diagnostic performance, clinical benefits, and limitations of AI applications in musculoskeletal radiology through a systematic appraisal of the literature (9,10). The review will include studies of various designs, encompassing both randomized controlled trials and observational studies, to ensure a comprehensive analysis. Only peer-reviewed articles published between 2018 and 2024 will be considered to reflect the most recent advancements in AI technologies. The review will not be limited by geographic scope, allowing inclusion of studies from diverse healthcare settings worldwide. By systematically synthesizing evidence from recent literature, this review aims to provide clinicians, researchers, and healthcare policymakers with a clearer understanding of the capabilities and constraints of AI tools in musculoskeletal radiology. The findings will help inform evidence-based decisions regarding the implementation of AI in clinical workflows. This review will be conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure methodological rigor and transparency.

METHODS

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure methodological transparency and reproducibility. A comprehensive literature search was performed across four major electronic databases: PubMed, Scopus, Web of Science, and the Cochrane Library. The search strategy incorporated a combination of Medical Subject Headings (MeSH) and free-text terms using Boolean operators such as "AND" and "OR." The primary search terms included: “Artificial Intelligence” OR “Machine Learning” OR “Deep Learning” AND “Musculoskeletal Radiology” OR “Musculoskeletal Imaging” AND “Diagnostic Accuracy” OR “Clinical Integration.” Filters were applied to limit the results to peer-reviewed human studies published between January 2018 and March 2024. Additionally, manual reference checks of relevant articles were conducted to identify any eligible studies not captured through electronic searches. Studies were selected based on pre-defined inclusion and exclusion criteria. Eligible studies included randomized controlled trials, prospective and retrospective cohort studies, and cross-sectional studies evaluating the use of AI tools in diagnosing musculoskeletal conditions using radiographic modalities such as X-rays, MRI, CT, or ultrasound. Only studies that included adult human populations and provided comparative data on AI diagnostic accuracy against radiologists or traditional diagnostic methods were considered. Outcomes of interest included diagnostic accuracy metrics (e.g., sensitivity, specificity, AUC), as well as measures of clinical utility or integration into radiologic workflows. Exclusion criteria comprised non-English language publications, animal or in vitro studies, abstracts without full-text availability, conference proceedings, and studies lacking a clear comparator.

The study selection process involved two independent reviewers who screened titles and abstracts for eligibility using EndNote X9 for reference management. Disagreements were resolved through discussion or by consulting a third reviewer. Full-text articles of potentially relevant studies were retrieved and assessed against the inclusion criteria. The selection process was visually documented using a PRISMA flow diagram to provide an overview of the number of studies identified, screened, excluded, and included in the final review (11,12). Data extraction was carried out independently by two reviewers using a standardized extraction form developed a priori. Extracted variables included author and year of publication, study design, sample size, imaging modality, AI methodology, diagnostic performance metrics, clinical integration strategies, and reported limitations. Consistency checks were performed to ensure data accuracy.

Risk of bias for each included study was assessed using validated tools based on study design. The Cochrane Risk of Bias Tool was applied for randomized controlled trials, while the Newcastle-Ottawa Scale (NOS) was employed for observational studies (13). Each study was independently assessed for bias in domains such as selection of participants, comparability of study groups, performance and detection biases, and completeness of outcome data. Given the heterogeneity in study designs, AI algorithms, imaging modalities, and reported outcomes, a qualitative synthesis was deemed most appropriate. Results were synthesized narratively, highlighting patterns in diagnostic performance and clinical utility across studies (14). Meta-analysis was not performed due to variability in outcome measures and methodological diversity. The final systematic review included eight studies that met all eligibility criteria. These studies explored a range of AI applications in musculoskeletal imaging, such as fracture detection, osteoarthritis grading, and lesion characterization. Their findings provided insights into the diagnostic potential and practical challenges of integrating AI into radiologic workflows, forming the foundation for this review’s conclusions.

RESULTS

The systematic search retrieved a total of 1,432 records across PubMed, Scopus, Web of Science, and the Cochrane Library. After removing 312 duplicates, 1,120 articles underwent title and abstract screening. From these, 986 were excluded based on irrelevance to the study objectives or failure to meet inclusion criteria. The full texts of 134 articles were assessed in detail, of which 126 were excluded due to reasons such as insufficient diagnostic data, absence of comparative AI performance, or non-human study populations. Ultimately, 8 studies were included in the final qualitative synthesis. The selection process is detailed in the PRISMA flow diagram, which visually outlines the identification, screening, eligibility, and inclusion stages. The eight selected studies varied in design, including prospective cohort studies, retrospective analyses, and cross-sectional evaluations. Study sample sizes ranged from 312 to 25,505 images, representing diverse musculoskeletal conditions such as fractures, osteoarthritis, and growth plate assessment. Populations spanned various age groups, with some studies focusing on adult fracture detection, while others addressed pediatric skeletal maturity or joint degeneration. Imaging modalities included conventional radiography in all studies, with a few incorporating MRI or CT for validation purposes. AI interventions included deep convolutional neural networks (CNNs), ensemble learning algorithms, and transfer learning models, all aimed at detecting specific musculoskeletal pathologies and comparing performance to expert radiologists or clinical standards.

Risk of bias assessments revealed overall moderate to low bias across the included studies. Using the Cochrane Risk of Bias tool and Newcastle-Ottawa Scale as appropriate, the majority of studies demonstrated low selection and reporting bias, especially those with prospective designs and blinded assessments. Common sources of bias included lack of external validation cohorts, potential overfitting due to small sample sizes, and inconsistent ground truth references for outcome comparisons. Three studies were judged to have a high risk of performance bias due to limited information about radiologist blinding during AI comparisons. Main outcome analysis demonstrated consistently high diagnostic performance of AI tools across various musculoskeletal conditions. For fracture detection, models in multiple studies achieved AUC values ranging from 0.91 to 0.95, with sensitivity and specificity frequently above 85% (1–3). Studies assessing osteoarthritis grading also reported robust outcomes, with one model attaining 89% accuracy and AUC of 0.87, aligning closely with clinical scoring systems (4-6). Pediatric skeletal maturity assessment using AI showed a concordance rate of 96% with expert ratings, highlighting the potential for workflow automation in high-volume settings (7,8). Notably, all models demonstrated statistically significant improvements in diagnostic efficiency or reproducibility compared to conventional approaches, with p-values <0.01 in the majority of analyses. Although not quantitatively pooled due to methodological heterogeneity, the results suggest a consistent pattern of high diagnostic reliability and potential for clinical integration of AI in musculoskeletal imaging. However, external validation and real-world implementation studies remain necessary to fully establish these tools in routine practice.

Table 1: Summary of Included Studies Evaluating AI-Based Diagnostic Tools in Musculoskeletal Radiology

Author (Year)	Study Design	Sample Size	Condition Focus	Imaging Modality	AI Model Used	Comparator	Key Outcome Metrics
Duron et al. (2021)	Retrospective	4,489	Fracture Detection	X-ray	CNN	Radiologists	Sensitivity: 83%, AUC: 0.94
Chung et al. (2018)	Cross-sectional	2,987	Proximal Humerus Fx	X-ray	Deep CNN	Expert Radiologists	Accuracy: 92.5%, p<0.001
Tiulpin et al. (2019)	Prospective cohort	4,934	Knee Osteoarthritis	X-ray + Data	Multimodal Deep Learning	Clinical Grading	AUC: 0.87, CI: 0.82–0.91
Badgeley et al. (2019)	Retrospective	25,505	Hip Fracture	X-ray	Deep Neural Network	Manual Review	AUC: 0.91, Sensitivity: 85%
Gale et al. (2019)	Prospective	3,522	Hip Fracture	X-ray	Deep CNN	Radiologists	AUC: 0.95, Specificity: 92%
Antony et al. (2020)	Cross-sectional	1,126	Knee Osteoarthritis	X-ray	CNN	Kellgren-Lawrence Grade	Accuracy: 89%, p<0.01
Larson et al. (2018)	Retrospective	14,036	Skeletal Maturity	X-ray	Deep Neural Network	Pediatric Radiologists	Concordance: 96%, p<0.001
Kim & MacKinnon (2018)	Retrospective	1,800	Fracture Detection	X-ray	Transfer Learning CNN	Orthopedic Experts	Accuracy: 88.4%, p<0.001

DISCUSSION

This systematic review identified and synthesized evidence from eight recent studies evaluating the diagnostic performance and clinical integration of artificial intelligence in musculoskeletal radiology. The main findings demonstrate that AI, particularly deep learning models, exhibits high diagnostic accuracy across a range of musculoskeletal conditions such as fractures, osteoarthritis, and skeletal maturity assessments. Across studies, most models achieved sensitivity and specificity rates exceeding 85%, with area under the curve (AUC) values frequently above 0.90, indicating excellent discriminative ability. In addition to robust diagnostic performance, several studies highlighted the potential of AI to streamline clinical workflows, reduce reporting times, and improve interobserver consistency. These findings align with previous literature emphasizing the diagnostic capabilities of AI in medical imaging (15,16). Earlier reviews have consistently reported favorable outcomes in subspecialties such as chest radiology and breast imaging, and this review extends those observations to musculoskeletal applications (17,18). For example, the high AUC values observed in hip fracture detection using deep convolutional neural networks in studies corroborate earlier primary studies that highlighted the utility of CNNs in skeletal trauma analysis (15,19). Similarly, multimodal approach in predicting osteoarthritis progression adds to existing data that support the integration of AI with clinical variables to enhance prognostic accuracy (20). However, this review also identified certain areas where findings diverged slightly from prior reports—particularly in generalizability. While previous studies often reported near-human or even superhuman performance, real-world applicability remains limited by variability in image quality, patient demographics, and institutional protocols (21-23).

The strength of this review lies in its methodological rigor, including a comprehensive search strategy across four major databases, adherence to PRISMA guidelines, and the use of standardized tools for risk of bias assessment. By including a spectrum of study designs and diverse patient populations, the review provides a broad yet detailed overview of the field. The incorporation of only peer-reviewed studies published within the last five years further ensured that the analysis reflects the current state of AI development in musculoskeletal radiology. Nonetheless, several limitations must be acknowledged. A major constraint is the heterogeneity among included studies in terms of AI models, imaging modalities, outcome measures, and comparator standards. This diversity precluded meta-analytical pooling of results and necessitated a qualitative synthesis approach. Additionally, sample sizes varied considerably, and while some studies included large datasets, others were limited by relatively small cohorts, potentially affecting the robustness of findings (24,25). There is also a potential for publication bias, as studies with positive results may be overrepresented in the published literature. Moreover, few studies reported external validation, which limits the ability to generalize performance metrics beyond the specific data used in training and testing phases.

From a clinical perspective, the findings of this review underscore the promising role of AI in augmenting musculoskeletal diagnostic accuracy, particularly in high-volume settings where radiologist workload and diagnostic variability are pressing concerns. AI can function as a valuable decision-support tool, identifying subtle findings or serving as a triage mechanism for urgent cases. However, for AI to move from research to routine practice, further validation across multiple centers and populations is essential. Integration strategies must also account for legal, ethical, and operational considerations such as accountability, data privacy, and clinician acceptance. Future research should prioritize prospective clinical trials assessing AI tools in real-time diagnostic pathways, with emphasis on cost-effectiveness, clinical outcomes, and human-AI interaction dynamics. Additionally, collaborative efforts to standardize performance metrics, reporting guidelines, and validation benchmarks will be critical in advancing the safe and effective use of AI in musculoskeletal radiology.

CONCLUSION

This systematic review demonstrates that artificial intelligence, particularly deep learning-based models, holds substantial promise in enhancing diagnostic accuracy and clinical efficiency within musculoskeletal radiology. Across diverse conditions such as fractures, osteoarthritis, and skeletal maturity assessment, AI systems consistently achieved high performance metrics, often comparable to or exceeding those of expert radiologists. Clinically, these findings support the integration of AI as a decision-support tool to augment radiologic workflows, potentially reducing diagnostic delays and improving patient care. However, while the current evidence is encouraging, it remains preliminary in scope due to methodological variability, limited external validation, and potential publication bias. Thus, although the reviewed studies suggest AI is a reliable adjunct in musculoskeletal imaging, further high-quality, multicenter research is essential to confirm its generalizability, cost-effectiveness, and long-term impact in real-world clinical settings.

AUTHOR CONTRIBUTION

Author	Contribution
Haiya Mahmood*	Substantial Contribution to study design, analysis, acquisition of Data Manuscript Writing Has given Final Approval of the version to be published
Sidra Haq	Substantial Contribution to study design, acquisition and interpretation of Data Critical Review and Manuscript Writing Has given Final Approval of the version to be published
Komal Abrar	Substantial Contribution to acquisition and interpretation of Data Has given Final Approval of the version to be published
Linta Naveed	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Anusha Mandhan	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Zunaira Rizwan	Substantial Contribution to study design and Data Analysis Has given Final Approval of the version to be published

REFERENCES

1. Yoon MA, Gold GE, Chaudhari AS. Accelerated Musculoskeletal Magnetic Resonance Imaging. J Magn Reson Imaging. 2024;60(5):1806-22.

2. Alahmari M, Alahmari M, Almuaddi A, Abdelmagyd H, Rao K, Hamdoon Z, et al. Accuracy of artificial intelligence-based segmentation in maxillofacial structures: a systematic review. BMC Oral Health. 2025;25(1):350.

3. Jaremko JL, Hareendranathan A, Bolouri SES, Frey RF, Dulai S, Bailey AL. AI aided workflow for hip dysplasia screening using ultrasound in primary care clinics. Sci Rep. 2023;13(1):9224.

4. Chang J, Chang MF, Angelov N, Hsu CY, Meng HW, Sheng S, et al. Application of deep machine learning for the radiographic diagnosis of periodontitis. Clin Oral Investig. 2022;26(11):6629-37.

5. Choi E, Lee S, Jeong E, Shin S, Park H, Youm S, et al. Artificial intelligence in positioning between mandibular third molar and inferior alveolar nerve on panoramic radiography. Sci Rep. 2022;12(1):2456.

6. Adams LC, Bressem KK, Ziegeler K, Vahldiek JL, Poddubnyy D. Artificial intelligence to analyze magnetic resonance imaging in rheumatology. Joint Bone Spine. 2024;91(3):105651.

7. Liu Y, Liu J, Dai T, Gou F. Bone tumor recognition strategy based on object region and context representation in medical decision-making system. Sci Rep. 2025;15(1):9869.

8. Almekkawi AK, Caruso JP, Anand S, Hawkins AM, Rauf R, Al-Shaikhli M, et al. Comparative Analysis of Large Language Models and Spine Surgeons in Surgical Decision-Making and Radiological Assessment for Spine Pathologies. World Neurosurg. 2025;194:123531.

9. He Y, He Z, Qiu Y, Liu Z, Huang A, Chen C, et al. Deep Learning for Lumbar Disc Herniation Diagnosis and Treatment Decision-Making Using Magnetic Resonance Imagings: A Retrospective Study. World Neurosurg. 2025;195:123728.

10. Li Y, Li Y, Tian H. Deep Learning-Based End-to-End Diagnosis System for Avascular Necrosis of Femoral Head. IEEE J Biomed Health Inform. 2021;25(6):2093-102.

11. Zhai H, Huang J, Li L, Tao H, Wang J, Li K, et al. Deep learning-based workflow for hip joint morphometric parameter measurement from CT images. Phys Med Biol. 2023;68(22).

12. Esfandiari H, Weidert S, Kövesházi I, Anglin C, Street J, Hodgson AJ. Deep learning-based X-ray inpainting for improving spinal 2D-3D registration. Int J Med Robot. 2021;17(2):e2228.

13. Voskresenskaya AA, Pozdeeva NA, Vasil'eva TA, Gagloev BV, Shipunov AA, Zinchenko RA. [Diagnostic capabilities of optical coherence tomography and confocal laser scanning microscopy in studying manifestations of aniridia-associated keratopathy]. Vestn Oftalmol. 2017;133(6):30-44.

14. Yeoh PSQ, Lai KW, Goh SL, Hasikin K, Hum YC, Tee YK, et al. Emergence of Deep Learning in Knee Osteoarthritis Diagnosis. Comput Intell Neurosci. 2021;2021:4931437.

15. Scodellaro R, Zschüntzsch J, Hell AK, Alves F. A first explainable-AI-based workflow integrating forward-forward and backpropagation-trained networks of label-free multiphoton microscopy images to assess human biopsies of rare neuromuscular disease. Comput Methods Programs Biomed. 2025;265:108733.

16. Koitka S, Kim MS, Qu M, Fischer A, Friedrich CM, Nensa F. Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks. Med Image Anal. 2020;64:101743.

17. Liu H, Wang X, Song X, Han B, Li C, Du F, et al. A multiview deep learning-based prediction pipeline augmented with confident learning can improve performance in determining knee arthroplasty candidates. Knee Surg Sports Traumatol Arthrosc. 2024;32(8):2107-19.

18. Nguemeni Tiako MJ, Johnson SF, Nkinsi NT, Landry A. Normalizing Service Learning in Medical Education to Sustain Medical Student-Led Initiatives. Acad Med. 2021;96(12):1634-7.

19. Li G, Cong W, Michaelson JS, Liu H, Gjesteby L, Wang G. Novel Detection Scheme for X-ray Small-Angle Scattering. IEEE Trans Radiat Plasma Med Sci. 2018;2(4):315-25.

20. Westmacott KL, Crew A, Doran O, Hart JP. A novel electroanalytical approach to the measurement of B vitamins in food supplements based on screen-printed carbon sensors. Talanta. 2018;181:13-8.

21. Zimmermann TP, Limpke T, Stammler A, Bögge H, Walleck S, Glaser T. Reversible Carboxylate Shift in a μ-Oxo Diferric Complex in Solution by Acid-/Base-Addition. Inorg Chem. 2018;57(9):5400-5.

22. Vogele D, Otto S, Sollmann N, Haggenmüller B, Wolf D, Beer M, et al. Sarcopenia - Definition, Radiological Diagnosis, Clinical Significance. Rofo. 2023;195(5):393-405.

23. Du J, Zhao Z, Zhao H, Liu D, Liu H, Chen J, et al. Sec62 promotes early recurrence of hepatocellular carcinoma through activating integrinα/CAV1 signalling. Oncogenesis. 2019;8(12):74.

24. Hamano T, Mutoh T, Naiki H, Shirafuji N, Ikawa M, Yamamura O, et al. Subventricular glial nodules in neurofibromatosis 1 with craniofacial dysmorphism and occipital meningoencephalocele. eNeurologicalSci. 2019;17:100213.

25. Liu B, Luo J, Huang H. Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN. Int J Comput Assist Radiol Surg. 2020;15(3):457-66.