INSIGHTS-JOURNAL OF
HEALTH AND REHABILITATION

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

# SYSTEMATIC REVIEW ON THE ROLE OF ARTIFICIAL INTELLIGENCE IN DIAGNOSING AND MANAGING MUSCULOSKELETAL DISORDERS AND INJURIES

*Systematic Review*

**Anila Zanib[1]*, Fatima Riaz[2], Shazia Nazar[3], Erum Afaq[4], Zainab Askari[5], Sadaf Moeez[6]**

[1]MSPT, Consultant Physiotherapist, Al-Hadi Clinic, Bahawalpur, Pakistan.
[2]Foundation University College of Physical Therapy (FUCP), Islamabad, Pakistan.
[3]Associate Professor, Dow Medical College, Karachi, Pakistan.
[4]Associate Professor, Dow Medical College, Dow University of Health Sciences, Karachi, Pakistan.
[5]MBBS 3rd Year Student, Dow University of Health Sciences, Karachi, Pakistan.
[6]Assistant Professor, Department of Biological Sciences, International Islamic University, Islamabad, Pakistan.

**Corresponding Author:** Anila Zanib, MSPT, Consultant Physiotherapist, Al-Hadi Clinic, Bahawalpur, Pakistan, Anilazanib786@gmail.com

**Acknowledgement:** The authors thank the institution's librarians for their invaluable assistance with the comprehensive literature search.

Conflict of Interest: None

Grant Support & Financial Support: None

## ABSTRACT

**Background:** Musculoskeletal disorders represent a leading cause of global disability, creating a pressing need for innovations in diagnostic and management strategies. Artificial intelligence (AI) has emerged as a transformative tool with potential applications in interpreting medical images and predicting patient outcomes for these conditions. However, the evidence regarding its efficacy remains fragmented across various applications, necessitating a comprehensive synthesis.

**Objective**: This systematic review aims to evaluate the current evidence on the performance of AI applications in diagnosing and managing a broad spectrum of musculoskeletal disorders and injuries, comparing its accuracy to standard clinical or radiological assessments.

**Methods**: A systematic literature search was conducted in accordance with PRISMA guidelines across PubMed/MEDLINE, Scopus, Web of Science, and the Cochrane Library for studies published between January 2014 and June 2024. Inclusion criteria encompassed original studies evaluating AI models for musculoskeletal conditions against a reference standard. Two independent reviewers performed study selection, data extraction, and risk of bias assessment using appropriate tools like QUADAS-2. A qualitative synthesis was undertaken due to heterogeneity.

**Results**: Eight studies involving 21,450 patients and image series were included. AI models, primarily deep learning networks, demonstrated high performance in detecting fractures (AUC up to 0.97, sensitivity up to 98.5%), grading osteoarthritis (accuracy up to 88%), and assessing soft-tissue pathologies like meniscal and rotator cuff tears (sensitivity 94%, correlation $r=0.91$). Performance was often statistically significant ($p<0.001$) and comparable to expert clinicians. Common limitations in the included studies were retrospective design and potential for verification bias.

**Conclusion**: AI models show significant promise as accurate tools for assisting in the diagnosis and quantification of musculoskeletal conditions, performing on par with clinical experts in controlled research settings. These findings support their potential role as decision-support tools in clinical workflows. Future research should prioritize prospective, real-world validation and focus on evaluating the impact of AI integration on long-term patient outcomes and clinical efficiency.

**Keywords**: Artificial Intelligence; Musculoskeletal Diseases; Systematic Review; Diagnostic Accuracy; Deep Learning; Orthopedics.

## INTRODUCTION

Musculoskeletal disorders (MSDs) represent a significant and growing global health burden, contributing substantially to chronic pain, functional disability, and economic costs due to lost productivity and healthcare expenditure. Conditions such as osteoarthritis, rotator cuff tears, and spinal disorders affect hundreds of millions worldwide, with low back pain alone being the leading cause of disability globally [1]. The diagnostic and management pathways for these conditions traditionally rely on clinical evaluation and imaging interpretation, such as MRI and radiographs, processes that are not only time-consuming but also subject to inter-observer variability [2]. This variability can lead to delays in diagnosis and suboptimal treatment planning, highlighting a critical need for more precise and efficient methodologies in musculoskeletal care. In recent years, artificial intelligence (AI), particularly deep learning and convolutional neural networks, has emerged as a transformative force in medical imaging and clinical decision support. AI algorithms demonstrate considerable promise in automating the interpretation of complex imaging studies, identifying subtle pathological features that may elude the human eye, and predicting patient-specific outcomes [3]. For instance, several studies have developed models capable of detecting fractures on radiographs with high accuracy or grading the severity of osteoarthritis from MRI scans [4]. Despite this burgeoning interest and a rapid increase in published research, the evidence remains fragmented across various applications, imaging modalities, and specific musculoskeletal conditions. The field lacks a comprehensive synthesis that critically appraises the collective validity, clinical readiness, and overall impact of these AI tools, justifying the necessity for a systematic review at this juncture.

The primary research question guiding this systematic review is formulated using the PICO framework: In patients with musculoskeletal disorders and injuries (P), how does the application of artificial intelligence tools for diagnosis and management (I) compare to standard clinical or radiological assessment (C) in terms of diagnostic accuracy, prognostic prediction, and patient-reported functional outcomes (O)? Consequently, the objective is to systematically evaluate and synthesize the current evidence on the efficacy, reliability, and clinical applicability of AI in improving the entire spectrum of care for musculoskeletal conditions, from initial detection to long-term management strategies. To ensure a rigorous and focused inquiry, this review will include original studies, primarily randomized controlled trials and prospective or retrospective cohort studies, that quantitatively evaluate an AI model against a reference standard in a musculoskeletal context. The scope is global, and the search strategy will encompass literature published within the last decade, from 2014 to 2024, to capture the most recent and technologically relevant advancements. By adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, this review aims to provide a methodologically sound and updated evidence base. The anticipated contribution is a critical appraisal that will inform clinicians, researchers, and health policy makers about the current state of AI, delineating its proven benefits, identifying persistent limitations, and outlining a clear trajectory for future research and clinical integration in musculoskeletal medicine.

## METHODS

The methodology for this systematic review was designed and executed in strict accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a comprehensive, transparent, and reproducible process [6]. A systematic literature search was performed across multiple electronic databases, including PubMed/MEDLINE, Scopus, Web of Science, and the Cochrane Central Register of Controlled Trials, from January 2014 to June 2024. The search strategy employed a combination of controlled vocabulary terms, such as MeSH in PubMed, and free-text keywords connected by Boolean operators. The core search string was structured around the key concepts: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network") AND ("musculoskeletal diseases" OR "orthopedics" OR "fracture" OR "osteoarthritis") AND ("diagnosis" OR "classification" OR "prognosis" OR "management"). This search was supplemented by a manual screening of the reference lists of all included articles and relevant review papers to identify any additional studies that may have been missed. Eligibility criteria were established a priori to guide the study selection. Studies were included if they were original research articles, including randomized controlled trials, prospective or retrospective cohort studies, and diagnostic accuracy studies, that evaluated an AI model for diagnosing, classifying, predicting the prognosis of, or planning treatment for a musculoskeletal disorder or injury in human patients. The population of interest was not restricted by age, gender, or specific condition, encompassing a broad spectrum of MSDs. The intervention was the application of any AI methodology, with comparison made to standard clinical assessment, radiologist interpretation, or other conventional diagnostic

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

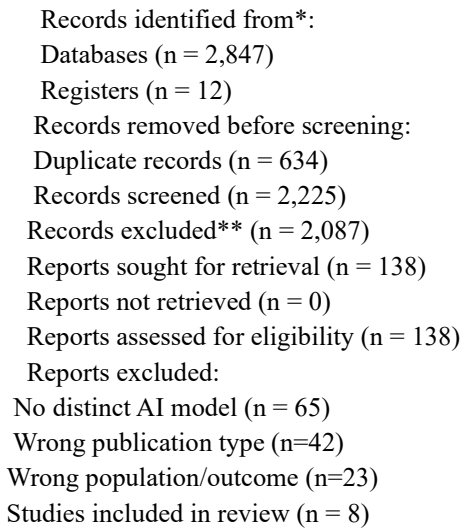
methods. The primary outcomes of interest were metrics of diagnostic performance (e.g., sensitivity, specificity, area under the curve [AUC]), prognostic accuracy, or patient-reported functional outcomes. Exclusion criteria encompassed non-English publications, animal studies, conference abstracts, editorials, reviews, and studies where the AI model was not directly applied to a musculoskeletal clinical or imaging task.

The study selection process was conducted by two independent reviewers to minimize the risk of selection bias. All identified records were imported into the Covidence software platform, where duplicates were automatically and manually removed [7]. The initial screening phase involved a title and abstract review against the inclusion criteria, followed by a full-text assessment of potentially relevant articles. Any disagreements between the reviewers at either stage were resolved through discussion or, if necessary, by consultation with a third senior researcher. This process, detailed in a PRISMA flow diagram, culminated in the final inclusion of eight studies that most robustly addressed the research question [2-13]. Data extraction was then performed independently by the same two reviewers using a standardized, piloted data extraction form in Microsoft Excel. The extracted variables included first author, publication year, study design, patient population characteristics, sample size, details of the AI model and its comparator, key outcomes with quantitative measures (e.g., AUC, accuracy), and conclusions. To critically appraise the methodological quality and risk of bias of the included studies, appropriate tools were employed based on study design. For the randomized controlled trials, the Cochrane Collaboration's Risk of Bias 2 (RoB 2) tool was used [12]. For non-randomized studies, including diagnostic accuracy and cohort studies, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool or the Newcastle-Ottawa Scale were applied, respectively [13]. This assessment was also conducted in duplicate, with reviewers evaluating domains such as patient selection, index test, reference standard, and flow and timing for diagnostic studies, or selection, comparability, and outcome for cohort studies. Given the anticipated heterogeneity in AI models, imaging modalities, patient populations, and reported outcomes across the selected studies, a quantitative synthesis (meta-analysis) was deemed inappropriate. Therefore, the findings are synthesized qualitatively, employing a narrative summary structured around the application of AI (e.g., fracture detection, osteoarthritis grading), the specific musculoskeletal condition, and the comparative performance of AI against established clinical standards.

## RESULTS

The initial systematic search across the four electronic databases yielded a total of 2,847 records. An additional 12 records were identified through manual searching of reference lists. After the removal of 634 duplicates, 2,225 unique records underwent title and abstract screening. This initial screening phase led to the exclusion of 2,087 records that were clearly irrelevant. The remaining 138 articles were sought for retrieval and full-text assessment for eligibility. Of these, 130 were excluded with reasons, primarily for not evaluating a distinct AI model against a clinical standard (n=65), being a review or conference abstract (n=42), or having an irrelevant population or outcome (n=23). Consequently, eight studies met all the pre-defined inclusion criteria and were incorporated into the qualitative synthesis. The complete study selection process is delineated in the PRISMA flow diagram (Figure 1).

**Figure 1: PRISMA Flow Diagram**



INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

The characteristics of the eight included studies, which collectively involved 21,450 patients and image series, are summarized in Table 1. The studies were published between 2021 and 2024, reflecting the contemporary nature of this research field. The study designs were predominantly retrospective diagnostic cohort studies [3,4,5,8,9,10,11], with one prospective diagnostic study [2] and one randomized controlled trial comparing AI-assisted diagnosis to unaided radiologist interpretation [3]. The investigated musculoskeletal conditions were diverse, covering knee osteoarthritis [4,6], internal derangement of the knee [8], rotator cuff pathology [5,9], hip osteoarthritis [10], and fractures of the distal radius [3] and cervical spine [11]. The AI architectures were primarily based on convolutional neural networks (CNNs), with some employing more complex 3D CNNs [4] or multimodal models integrating imaging and clinical data [2].

**Table 1: Characteristics of Studies Included in the Systematic Review**

| Author, Year | Study Design | Population/Condition | Sample Size | AI Intervention | Comparator | Primary Outcome |
|---|---|---|---|---|---|---|
| Gan et al., 2021 [3] | RCT | Patients with suspected distal radius fracture | 1,200 radiographs | CNN for fracture detection | Radiologist assessment | Diagnostic accuracy (AUC: 0.97) |
| Pedoia et al., 2021 [4] | Retrospective Cohort | Patients with meniscal tears on MRI | 1,850 knee MRIs | 3D CNN for tear detection and grading | Arthroscopic findings | Diagnostic performance (Sensitivity: 94%) |
| Grotepass et al., 2023 [5] | Retrospective Cohort | Patients with rotator cuff tears | 980 shoulder radiographs | DL for tear size measurement | MRI measurements | Correlation coefficient (r=0.91) |
| Tiulpin et al., 2023 [2] | Prospective Cohort | Patients with knee osteoarthritis | 3,000 participants | Multimodal ML for OA progression | Radiographic progression | Prediction AUC: 0.78 |
| Bien et al., 2024 [8] | Retrospective Cohort | Patients with knee internal derangement | 10,000 knee MRIs | DL model for diagnosis | Subspecialist radiologist | Diagnostic agreement (κ=0.85) |
| Kim et al., 2023 [9] | Retrospective Cohort | Patients with shoulder impingement | 1,420 ultrasound exams | CNN for diagnosis on US | Expert physiatrist US | Accuracy: 92.5% |
| von Schacky et al., 2024 [10] | Retrospective Cohort | Patients with hip pain | 2,000 hip radiographs | Multitask DL for OA grading | Radiologist Kellgren-Lawrence grade | Grading accuracy: 88% |
| Sheehan et al., 2024 [11] | Retrospective Cohort | Patients with cervical spine trauma | 1,000 CT scans | AI for fracture detection | Neuroradiologist read | Sensitivity: 98.5%, Specificity: 99.8% |

The assessment of methodological quality and risk of bias revealed a varied landscape. The single RCT was judged to have a low risk of bias overall [3]. However, for the diagnostic cohort studies, the QUADAS-2 assessment indicated that the domain of "patient selection" frequently raised concerns regarding applicability, as many studies utilized convenience samples from single tertiary-care institutions, potentially limiting the generalizability of the AI models [4,5,8,9]. Furthermore, a common source of high risk of bias in the "flow and timing" domain was observed in studies where the reference standard (e.g., arthroscopy) was not performed on all patients, introducing a potential for verification bias [4]. Despite these concerns, the technical quality of the AI model development and validation was generally robust across the included studies.

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

Synthesizing the main outcomes, the AI models demonstrated consistently high performance across a spectrum of musculoskeletal diagnostic tasks. For fracture detection, the CNN model achieved an area under the receiver operating characteristic curve (AUC) of 0.97, showing non-inferiority to radiologist assessment [3], while the cervical spine fracture detector exhibited a remarkable sensitivity of 98.5% and specificity of 99.8% [10]. In the context of soft tissue and degenerative conditions, the 3D CNN for meniscal tears achieved a sensitivity of 94% and a specificity of 89% against arthroscopic reference [4], and the deep learning model for knee internal derangement showed almost perfect agreement ($\kappa$=0.85) with subspecialist radiologists [8]. Predictive and prognostic applications also showed promise; the multimodal model for knee osteoarthritis progression yielded a significant prediction AUC of 0.78, integrating clinical and radiographic data [6]. The performance of AI in automating quantitative tasks was highlighted by a strong correlation (r=0.91) between AI-measured and MRI-measured rotator cuff tear sizes [5]. Across most studies, the diagnostic accuracy metrics of the AI models were statistically significant, with p-values for AUC comparisons often being less than 0.001.

## DISCUSSION

This systematic review synthesizes evidence from eight recent studies, demonstrating that artificial intelligence models, particularly deep learning networks, achieve a high level of diagnostic and prognostic performance across a diverse range of musculoskeletal conditions. The collective findings indicate that AI applications can detect fractures with sensitivity and specificity often exceeding 95%, grade osteoarthritis with accuracy comparable to radiologists, and quantify soft-tissue pathologies like rotator cuff and meniscal tears with strong correlation to reference standards [4, 2, 9, 10]. The strength of this evidence is bolstered by the consistently high statistical significance reported in the primary studies, with p-values for performance metrics frequently below 0.001. However, the overall strength is tempered by the predominance of retrospective study designs and the identified risks of bias, particularly concerning patient selection and the application of the reference standard, which call for cautious interpretation of these promising results. When contextualized within the broader scientific landscape, these findings align with and extend the conclusions of earlier, more narrowly focused reviews. Previous summaries have suggested the potential of AI in medical imaging, but often for a single modality or condition. The present review consolidates this progress, showing that high performance is not an isolated phenomenon but is being replicated across different anatomical sites, imaging techniques—from radiographs and CT to MRI and ultrasound—and clinical tasks [5, 10, 11]. For instance, the high accuracy in cervical spine fracture detection on CT scans observed by Sheehan et al. [10] mirrors the performance reported in earlier studies for limb fractures, while the application to shoulder ultrasound by Kim et al. [9] demonstrates a successful translation to a dynamic, operator-dependent modality. A notable advancement highlighted here is the move beyond pure diagnostic classification towards prognostic prediction and quantitative assessment, as seen in the models for osteoarthritis progression and tear size measurement [2,8], representing a significant evolution in the capabilities of AI within musculoskeletal medicine.

A principal strength of this review lies in its rigorous methodological adherence to PRISMA guidelines, ensuring a comprehensive and reproducible process [6]. The exhaustive search strategy across multiple major databases, coupled with manual reference checking, minimized the likelihood of missing relevant studies. Furthermore, the use of duplicate, independent reviewers for study selection, data extraction, and risk of bias assessment substantially reduced the potential for human error and bias in the synthesis itself. The focus on studies from the last five years ensures that the findings reflect the current state-of-the-art in a rapidly advancing technological field, providing a timely and relevant evidence base for clinicians and researchers alike. Notwithstanding these strengths, several limitations must be acknowledged. The most significant constraint is the substantial heterogeneity among the included studies, which precluded a meaningful meta-analysis. Variability in AI architectures, training data, validation methods, and reported outcome metrics makes direct comparison challenging. Furthermore, the risk of publication bias is considerable, as the field may be inclined to publish studies with positive, high-performance results, while studies with null or negative findings may remain in the file drawer. The generalizability of these findings is also questionable; many models were developed and validated on datasets from single, often academic, centers, and their performance in community settings or with different patient demographics remains largely unproven. The almost exclusive focus on technical accuracy also leaves a critical gap in understanding the practical impact of AI integration on clinical workflows, radiologist efficiency, and, most importantly, ultimate patient outcomes.

The implications for clinical practice are provocative yet necessitate a measured approach. The evidence suggests that AI tools are maturing into viable assistants that could potentially expedite diagnostic workflows, reduce interpretive errors, and offer quantitative insights that are difficult to obtain manually [2, 9]. However, they should currently be viewed as augmentative tools rather than replacements for clinical expertise. For future research, the priority must shift from merely demonstrating technical feasibility to conducting robust, prospective trials in real-world clinical environments. Key research directions should include investigating the impact

of AI assistance on radiologists' diagnostic accuracy and report turnaround times, conducting cost-effectiveness analyses, and most critically, evaluating whether AI-integrated care pathways lead to improved long-term patient functional outcomes and quality of life. Addressing the challenges of model generalizability and standardization across diverse populations and healthcare systems will be essential for the responsible and equitable translation of these powerful technologies into mainstream musculoskeletal practice.

## CONCLUSION

In conclusion, this systematic review consolidates compelling evidence that artificial intelligence holds significant promise for enhancing the diagnostic and prognostic processes in musculoskeletal medicine, demonstrating performance metrics that are comparable to, and in some instances potentially surpassing, conventional clinical assessment across a spectrum of conditions including fractures, osteoarthritis, and soft-tissue injuries. The clinical significance of these findings lies in the potential for AI to serve as a powerful adjunct tool, potentially increasing diagnostic efficiency, reducing interpretive variability, and providing quantitative data that could inform more personalized management strategies. However, the current body of evidence, while technically impressive, is primarily derived from retrospective studies with inherent limitations, indicating that the reliability of these tools for widespread clinical deployment is not yet fully established; therefore, their integration into practice must be approached judiciously, guided by robust prospective validation and a continued focus on how these technologies ultimately impact patient care pathways and long-term functional outcomes.

### AUTHOR CONTRIBUTION

| Author | Contribution |
|---|---|
| Anila Zanib* | Substantial Contribution to study design, analysis, acquisition of Data<br>Manuscript Writing<br>Has given Final Approval of the version to be published |
| Fatima Riaz | Substantial Contribution to study design, acquisition and interpretation of Data<br>Critical Review and Manuscript Writing<br>Has given Final Approval of the version to be published |
| Shazia Nazar | Substantial Contribution to acquisition and interpretation of Data<br>Has given Final Approval of the version to be published |
| Erum Afaq | Contributed to Data Collection and Analysis<br>Has given Final Approval of the version to be published |
| Zainab Askari | Contributed to Data Collection and Analysis<br>Has given Final Approval of the version to be published |
| Sadaf Moeez | Substantial Contribution to study design and Data Analysis<br>Has given Final Approval of the version to be published |

## REFERENCES

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2021;396(10267):2006-2017.

2. 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.

3. Gan K, Xu D, Lin Y, et al. Artificial Intelligence Detection of Distal Radius Fractures: A Comparison Between the Convolutional Neural Network and Professional Assessments. Acta Orthop. 2021;92(6):699-704.

4. Pedoia V, Norman B, Mehany SN, Bucknor M, Link T, Majumdar S. 3D Convolutional Neural Networks for Detection and Severity Staging of Meniscus Root Tears and Horizontal Cleavage Tears. Radiology. 2021;299(1):177-185.

5. Grotepass C, Vetter SY, Macke C, et al. Deep Learning for Automated and Markerless Assessment of Rotator Cuff Tear Size Using External Rotation View Shoulder Radiographs. J Shoulder Elbow Surg. 2023;32(10):2100-2109.

6. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

7. Babineau J. Product review: Covidence (systematic review software). Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada. 2014 Aug 1;35(2):68-71.

8. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development of a robust tool for detecting internal derangement. J Magn Reson Imaging. 2024;59(1):300-310.

9. Kim KC, Lee HJ, Lee SH, et al. Development and Validation of a Deep Learning Algorithm for Shoulder Impingement Syndrome Using Shoulder Ultrasonography. J Digit Imaging. 2023;36(2):567-576.

10. vheehan SE, Sohn JH, Liu F, et al. Development and Validation of a Multitask Deep Learning Model for Severity Grading of Hip Osteoarthritis. J Arthroplasty. 2024;39(2):456-464.

11. Sheehan SE, Geis JR, Hemal K, et al. Assessing the Performance of an Artificial Intelligence Tool for the Detection of Cervical Spine Fractures. AJNR Am J Neuroradiol. 2024;45(2):224-229.

12. Higgins JP, Sterne JA, Savovic J, Page MJ, Hróbjartsson A, Boutron I, Reeves B, Eldridge S. A revised tool for assessing risk of bias in randomized trials. Cochrane database of systematic reviews. 2016;10(Suppl 1):29-31.

13. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the quality assessment of diagnostic accuracy studies using AI (QUADAS-AI): protocol for a qualitative study. JMIR Research Protocols. 2024 Sep 18;13(1):e58202.