Research Letter

Can Artificial Intelligence Diagnose Knee Osteoarthritis?

Mihir Tandon[1], BA; Nitin Chetla[2], BS; Adarsh Mallepally[3]; Botan Zebari[4], BS; Sai Samayamanthula[2], BA; Jonathan Silva[1], BS; Swapna Vaja[5], BS; John Chen[1], BS; Matthew Cullen[1], BS; Kunal Sukhija[6], MD

[1]Albany Medical College, Albany, NY, United States
[2]University of Virginia School of Medicine, Charlottesville, VA, United States
[3]School of Medicine, Virginia Commonwealth University, Richmond, VA, United States
[4]St. James School of Medicine, Binghamton, NY, United States
[5]Rush Medical College, Chicago, IL, United States
[6]Kaweah Health, Visalia, CA, United States

Corresponding Author:
Mihir Tandon, BA
Albany Medical College
43 New Scotland Ave
Albany, NY, 12208
United States
Phone: 1 3322488708
Email: tandonm@amc.edu

Related Article:
This is a corrected version. See correction statement in: https://biomedeng.jmir.org/2025/1/e82980

Abstract

This study analyzed the capability of GPT-4o to properly identify knee osteoarthritis and found that the model had good sensitivity but poor specificity in identifying knee osteoarthritis; patients and clinicians should practice caution when using GPT-4o for image analysis in knee osteoarthritis.

(JMIR Biomed Eng 2025;10:e67481) doi: 10.2196/67481

KEYWORDS

large language model; ChatGPT; GPT-4o; radiology; osteoarthritis; machine learning; X-rays; osteoarthritis detection

Introduction

Osteoarthritis often affects the knee, causing pain and disability, and is typically diagnosed by X-ray [1]. Advancements in artificial intelligence (AI) offer potential to automate image analysis, reducing diagnostic burden [2]. Given its widespread availability, tools like ChatGPT have potential as point-of-care diagnostic aids. AI has already been incorporated on the physician side through clinical decision support systems and robotic surgery. On the patient side, AI is used in applications such as virtual health assistants [3].

Orthopedic surgeons, radiologists, and primary care physicians can use AI tools to streamline their workflows and reduce errors while analyzing imaging for pathologies like osteoarthritis. Moreover, patients use ChatGPT to analyze their imaging to further understand their condition [4]. The ability of AI to read other radiological images (eg, computed tomography angiograms) has been shown to be subpar [5]. However, studies have shown that AI can perform well with X-rays [6]. As such, it is increasingly important for physicians to understand AI's strengths and limitations to assess its use in imaging and guide patients using AI for self-diagnosis.

Methods

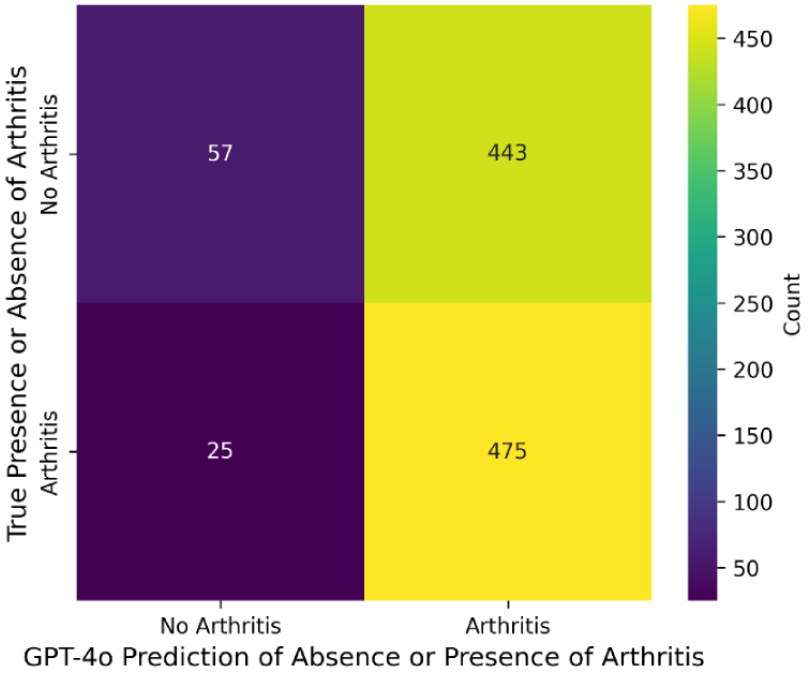
We queried ChatGPT (using the GPT-4o version) and assessed its performance in classifying 500 X-ray images of normal knees and 500 images of knees with osteoarthritis from a publicly available Kaggle database [7]. Images were verified based on consensus among radiologists. A single standardized prompt was used: "This is an x-ray image found on examination, the multiple-choice question is as follows. Based on the x-ray image, does the patient have A) no osteoarthritis, B) osteoarthritis." Key metrics included accuracy, sensitivity, and specificity. No images were rejected by ChatGPT. The code used for statistical analysis is included in Multimedia Appendix 1.

Results

The model's performance in distinguishing osteoarthritis from nonosteoarthritis knee X-rays was mixed. The high recall (0.950, 95% CI 0.964-0.943) suggests that the model was sensitive in identifying arthritis cases, while the low specificity (0.114, 95% CI 0.134-0.104) indicated a poor ability to correctly identify nonosteoarthritis cases. The F_1-score (0.670, 95% CI 0.699-0.655) balanced precision and recall, showing moderate effectiveness, but the precision (0.517, 95% CI 0.548-0.501) reflected that about half the predicted osteoarthritis cases were correct. Accuracy was 0.532 (95% CI 0.563-0.516). Figure 1 shows sensitivity and specificity.

The binomial test, where the null hypothesis assumed the model's accuracy was 50% or less, indicated that the model was statistically better than random chance (P=.02). Additionally, the χ^2 test (P<.001) indicated a strong dependence between the model's predictions and the actual labels, demonstrating that its classifications were not purely random. However, the significance of this test should be interpreted with caution, as it does not necessarily reflect high accuracy or clinical reliability.

Figure 1. Sensitivity and specificity of Chat-GPT4o in analyzing knee osteoarthritis X-rays.



Discussion

The model had difficulty distinguishing between "not arthritis" and "arthritis." While the recall for arthritis was high (0.950), indicating strong performance in identifying true arthritis cases, the low specificity (0.114) reflects a significant number of false positives, with many nonarthritis cases misclassified as arthritis. This bias toward predicting arthritis lowered precision (0.517) and accuracy (0.532); similar misclassification issues have been reported in other ChatGPT studies [8].

Limitations include, first, that the prompt was binary. A binary prompt was used because it would have been difficult to analyze data obtained with an open-ended prompt. Second, the dataset was small; a larger dataset would have yielded more robust conclusions.

Even with its limitations, this study presents important data on GPT4o's use in imaging for diagnosing osteoarthritis. This is vital, as our understanding of tools like this in health care contexts is limited. These results suggest a need for better class balance and improved feature differentiation. Similar misclassification patterns have been noted in previous studies, where overlapping features led to false positives [9]. A higher-resolution, more comprehensively annotated osteoarthritis dataset could improve model training, enhancing overall accuracy, sensitivity, and specificity. Thus, future work should focus on analyzing larger datasets and refining the model to handle more nuanced cases more effectively, improving performance statistics. Using image preprocessing techniques, such as contrast enhancement and noise reduction, and including metadata like medical history and clinical presentation could also help distinguish osteoarthritis from anatomical variations.

Our results suggest that clinicians should use ChatGPT cautiously and as a screening tool prior to their own validation to help mitigate misclassification. Clinicians should also educate patients about the risks of using AI for self-diagnosis of osteoarthritis based on X-rays. Despite its shortcomings, AI has potential for developing more reliable diagnostic models for osteoarthritis.

Authors' Contributions

Conceptualization: NC (lead), MT (equal), KS (equal)
Data curation: AM (lead), MT (equal), SS (supporting), SV (supporting), JC (supporting)
Formal analysis: JC (lead), JS (supporting), MC (supporting), SV (supporting), AM (supporting)
Funding acquisition: KS (lead)
Investigation: SS (lead), KS (equal), BZ (supporting), SV (supporting)
Methodology: MT (lead), NC (equal), KS (equal), AM (supporting)
Resources: SV (lead), JC (supporting)
Software: JC (lead), AM (supporting)
Supervision: KS (lead), MT (equal), NC (equal)
Validation: JS (lead), JC (equal), MC (equal)
Visualization: MT (lead), MC (equal), SS (supporting)
Writing – original draft: MT (lead), NC (equal), BZ (supporting), SS (supporting), AM (supporting)
Writing – review & editing: JS (lead), SV (equal), JC (equal), MC (supporting), KS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Code for analysis and prompting.
[DOCX File , 17 KB-Multimedia Appendix 1]

References

1. Choi MS, Lee DK. The effect of knee joint traction therapy on pain, physical function, and depression in patients with degenerative arthritis. J Kor Phys Ther. Oct 31, 2019;31(5):317-321. [FREE Full text] [doi: 10.18857/jkpt.2019.31.5.317]
2. Bejarano A. The benefits of artificial intelligence in radiology: transforming healthcare through enhanced diagnostics and workflow efficiency. Rev Contemp Sci Acad Stud. Aug 30, 2023;3(8):1-4. [doi: 10.55454/rcsas.3.08.2023.005]
3. Chatterjee I, Ghosh R, Sarkar S, Das K, Kundu M. Revolutionizing innovations and impact of artificial intelligence in healthcare. Int J Multidiscip Res. May 14, 2024;6(3):19333. [doi: 10.36948/ijfmr.2024.v06i03.19333]
4. Zhang Z, Citardi D, Wang D, Genc Y, Shan J, Fan X. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. Health Informatics J. 2021;27(2):14604582211011215. [FREE Full text] [doi: 10.1177/14604582211011215] [Medline: 33913359]
5. Young A, Tan K, Tariq F, Jin MX, Bluestone AY. Rogue AI: cautionary cases in neuroradiology and what we can learn from them. Cureus. Mar 2024;16(3):e56317. [FREE Full text] [doi: 10.7759/cureus.56317] [Medline: 38628986]
6. Wu JT, Wong KCL, Gur Y, Ansari N, Karargyris A, Sharma A, et al. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. JAMA Netw Open. Oct 01, 2020;3(10):e2022779. [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.22779] [Medline: 33034642]
7. Kabir F. Osteoarthritis prediction. Kaggle. URL: https://www.kaggle.com/datasets/farjanakabirsamanta/osteoarthritis-prediction [accessed 2024-09-01]
8. Dalalah D, Dalalah OM. The false positives and false negatives of generative AI detection tools in education and academic research: the case of ChatGPT. Int J Manag Educ. Jul 2023;21(2):100822. [doi: 10.1016/j.ijme.2023.100822]
9. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep. Dec 17, 2023;13(1):20159. [FREE Full text] [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]

Abbreviations

AI: artificial intelligence

Edited by S Rizvi, T Leung; submitted 12.10.24; peer-reviewed by Y Chaibi, A Jahnen, M Nayak; comments to author 25.02.25; revised version received 13.03.25; accepted 25.03.25; published 23.04.25

Please cite as:
Tandon M, Chetla N, Mallepally A, Zebari B, Samayamanthula S, Silva J, Vaja S, Chen J, Cullen M, Sukhija K
Can Artificial Intelligence Diagnose Knee Osteoarthritis?
JMIR Biomed Eng 2025;10:e67481
URL: https://biomedeng.jmir.org/2025/1/e67481
doi: 10.2196/67481
PMID: 40266670