Evaluating the Performance of ChatGPT-4o in Classifying Knee X-rays for Osteoarthritis Detection: Challenges in Sensitivity and Specificity

Mihir Tandon, Nitin Chetla, Ardash Mallepally, Bo Zebari, Sai Samayamanthula, Kunal Sukhija

Submitted to: JMIR AI
on: October 12, 2024

Table of Contents

Evaluating the Performance of ChatGPT-4o in Classifying Knee X-rays for Osteoarthritis Detection: Challenges in Sensitivity and Specificity

Mihir Tandon[1] BA; Nitin Chetla[2] BS; Ardash Mallepally[3] BS; Bo Zebari[4] BS; Sai Samayamanthula[2] BS; Kunal Sukhija[5] MD

[1]Albany Medical College Albany US
[2]University of Virginia School of Medicine Charlottesville US
[3]Virginia Commonwealth University School of Medicine Richmond US
[4]State University of New York, Binghamton Binghamton US
[5]Kaweah Health Visalia US

Corresponding Author:
Mihir Tandon BA
Albany Medical College
43 New Scotland Ave
Albany
US

Abstract

Large language models have gained popularity in healthcare in multiple fields. One of these fields is radiology. Patients may use tools like Chat-GPT4o to scan their imaging to better understand their pathology. Clinicians may also use Chat-GPT4o to increase productivity and reduce human error. However, given this is a new technology, we do not know the diagnostic efficacy of Chat-GPT4o in the field of radiology. The aim of this study was to analyze the capability of Chat-GPT4o in properly identifying knee osteoarthritis.

One thousand x-rays were given to Chat-GPT. Five hundred were normal knee x-rays, and the others were knees with osteoarthritis, vetted by radiologists. The x-rays were provided from an online publicly available database on Kaggle. Chat-GPT4o had good sensitivity but poor specificity in identifying knee osteoarthritis. It had a high level of false positives and poor precision.

Overall, patients and clinicians should practice caution when using Chat-GPT4o to analyze imaging in knee osteoarthritis.

(JMIR Preprints 12/10/2024:67481)
DOI: https://doi.org/10.2196/preprints.67481

Original Manuscript

Introduction

Osteoarthritis is a degenerative joint disease that commonly affects the knee, leading to pain, functional limitations, and decreased quality of life.[1] Advancements in artificial intelligence offer the potential to automate image analysis, reducing diagnostic burden.[2] Given its widespread availability, tools like ChatGPT-4o have the potential to be used as point-of-care diagnostic aids. In healthcare, artificial intelligence has already been incorporated on the provider side through avenues such as clinical decision support systems and robotic surgery. On the patient side, artificial intelligence is used in forms like virtual health assistants.[3]

Orthopaedic surgeons, radiologists, and primary care physicians may be able to use these tools to streamline workflow and catch human error while analyzing imaging for pathologies like osteoarthritis. Moreover, patients have access to their imaging, which allows them to autonomously use Chat-GPT4o to analyze their imaging to further understand their conditions.[4] As such, it is increasingly important for physicians to realize both the capabilities and limitations of artificial intelligence tools to better comprehend the efficacy of these tools for their own use in imaging and to interact with and educate patients who may be using artificial intelligence tools for self-diagnosis.

Methods

ChatGPT-4o was queried, assessing its performance in classifying 500 x-ray images each of normal knees and knees with osteoarthritis from a publicly available Kaggle database.[5] Images were verified using clinical examination, radiologist consensus, or follow-up imaging. A single standardized prompt was used:

Prompt: 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 calculated include accuracy, sensitivity, and specificity. Images that ChatGPT refused to answer were excluded.

Results

The results indicate that the model's performance in distinguishing osteoarthritis from non-osteoarthritis knee x-rays is mixed. The high recall (0.95, CI: 0.964-0.943) suggests that the model is very sensitive to identifying arthritis cases, while the low specificity (0.114, CI: 0.134-0.104) indicates a poor ability to correctly identify non-osteoarthritis cases. The F1 score (0.67, CI: 0.699-0.655) balances precision and recall, showing moderate overall effectiveness, but the precision (0.517, CI: 0.548-0.501) reflects that only about half of the predicted osteoarthritis cases were correct.

Discussion

The model had difficulty distinguishing between "not arthritis" and "arthritis." While the recall for arthritis was high (0.95), indicating strong performance in identifying true arthritis cases, the low specificity (0.114) reflects a significant number of false positives, with many non-arthritis cases being misclassified as arthritis. This bias toward predicting arthritis lowered overall precision (0.517) and accuracy (0.532).

There are a few limitations to this study. First, the specific prompt given to Chat-GPT4o was binary. The reason a binary prompt was used for this study was because it would be difficult to analyze data in the case of an open-ended prompt. Next, the dataset may be limited, with only 500 images. With a greater dataset size, the conclusions drawn would be more robust. Finally, exclusion of images in which Chat-GPT4o refused to respond reduces the ability of this work to analyze the wide variety of x-rays present in patients. However, no x-rays were excluded in this study.

Even with such limitations, this study presents important data on Chat-GPT4o's use in imaging related to diagnosing osteoarthritis. This is vital as our understanding of these tools in healthcare 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.[6] Future work should focus on expanding the dataset and refining the model to handle ambiguous cases more effectively. Based on these results, clinicians should use Chat-GPT4o cautiously for their own use and caution patients using artificial intelligence for self-diagnosis of osteoarthritis based on X-rays they have. However, this is still great potential in the use of these tools.

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 Korean Phys Ther*. 2019;31(5):317-321. 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*. 2023;3(8). doi:10.55454/rcsas.3.08.2023.005

3. - IC, - RG, - SS, - KD, - MK. Revolutionizing Innovations and Impact of Artificial Intelligence in

Healthcare. *Int J Multidiscip Res*. 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. doi:10.1177/14604582211011215

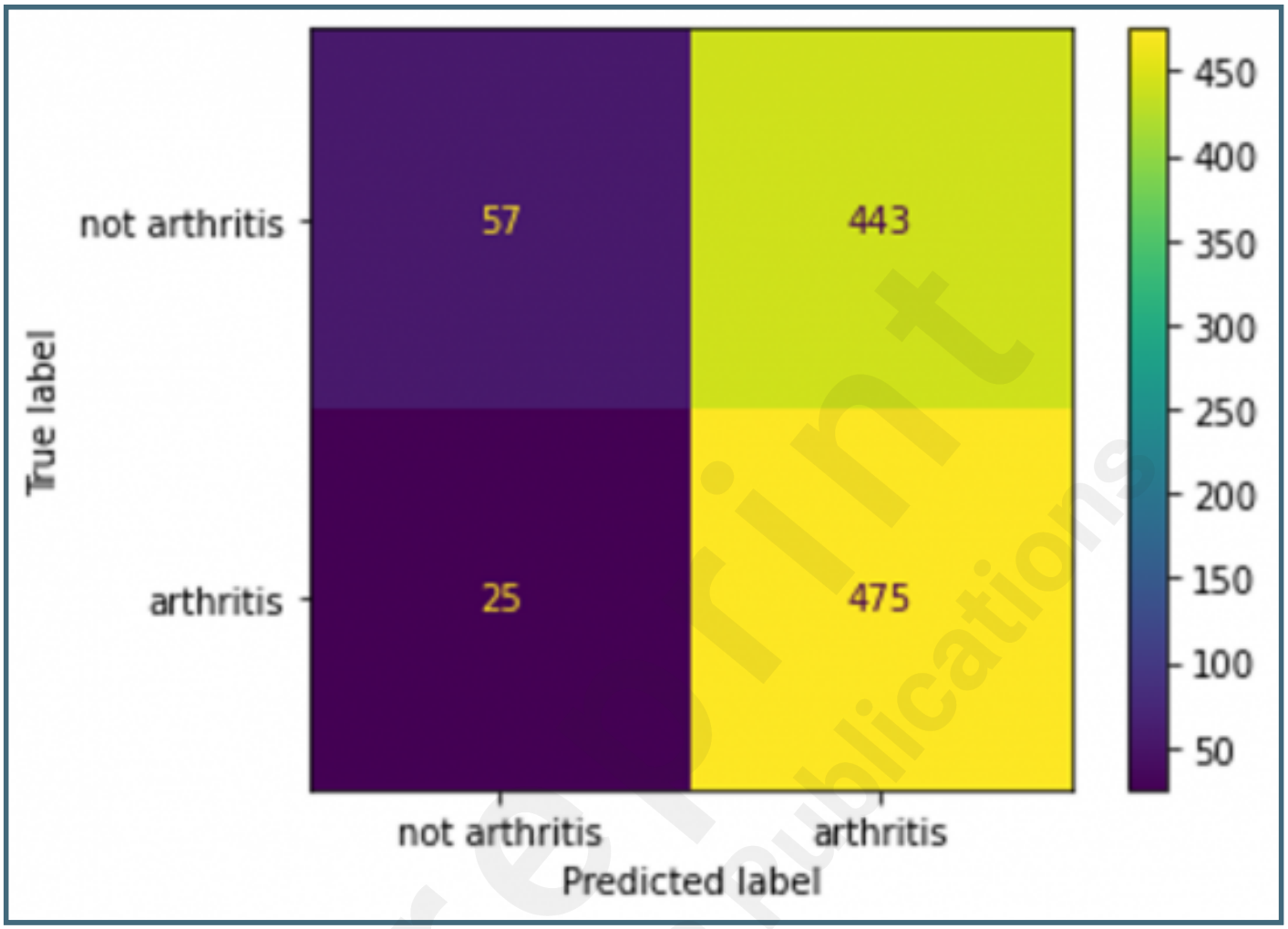
5. Kabir F. Osteoarthritis Prediction. Accessed September 1, 2024. https://www.kaggle.com/datasets/farjanakabirsamanta/osteoarthritis-prediction

6. Truhn D, Weber CD, Braun BJ, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci Rep*. 2023;13(1):20159. doi:10.1038/s41598-023-47500-2

Supplementary Files

Figures

Confusion matrix of Chat-GPT4o in analyzing knee osteoarthritis x-rays.



Statistics on diagnostic efficacy of Chat-GPT 4o in analyzing knee osteoarthritis X-rays.

Accuracy	F1	Recall	Precision	Specificity
0.532 (95% CI: 0.563-0.516)	0.67 (95% CI: 0.699-0.655)	0.95 (95% CI: 0.964-0.943)	0.517 (95% CI: 0.548-0.501)	0.114 (95% CI: 0.134-0.104)