İZMİR DEMOKRASİ ÜNİVERSİTESİ

IMCIDU 2024

CONGRESS BOOKLET

FULL TEXTS
&
ABSTRACTS

6TH INTERNATIONAL MEDICAL CONGRESS OF IZMIR DEMOCRACY UNIVERSITY

IZMIR, TURKIYE

05-06 DECEMBER 2024

https://imcidu.idu.edu.tr/

imcidu@idu.edu.tr

ID 568

The Ability of AI-Based Chatbots to Interpret Mammography Images: A Comparison Between Chat-GPT 4o and Claude 3.5

Mahmut Altuğ Altın, Betül Nalan Karahan, Emre Emekli

mahmutaltugaltin@gmail.com
COAN: 0027IMCIDU2024

ABSTRACT

Background: AI-based chatbots are widely used today. They have been researched in various fields within medicine, such as evaluating responses to patient questions, assessing answers on medical exams, question writing, and scientific article drafting (1-3). With recent updates, these chatbots are now able to analyze and interpret images. This study aims to evaluate mammography image interpretation using Chat-GPT 4o and Claude 3.5 chatbots.

Materials and Methods: Fifty-three mammography images, reported by consensus between two radiologists, were included in the study. These images were distributed as 10 images each for BI-RADS categories 0, 1, 2, 4, and 5, and 3 images for BI-RADS-3. Each mammography image consisted of one craniocaudal (CC) and one mediolateral oblique (MLO) standard view. Both views of each mammogram were presented to the chatbots in new tabs, asking for BI-RADS classification and breast typing. For Chat-GPT 4o, this process was repeated after one day. Subsequently, accuracy rates for breast parenchymal type and BI-RADS classification were calculated for both Chat-GPT 4o and Claude 3.5. Additionally, BI-RADS 1 and 2 were evaluated as benign categories, while BI-RADS 4 and 5 were considered malignant; since no change in patient management occurred for these categories, a single accuracy calculation was performed. Intra-observer agreement was also assessed for Chat-GPT 4o.

Results: For BI-RADS classification, the accuracy of Chat-GPT 4o was 18.87% on the first attempt and 26.42% on the second; for Claude 3.5, it was 18.87%. When BI-RADS 1 and 2 were considered benign, and BI-RADS 4 and 5 malignant, the overall accuracy for these 40 patients was calculated as 57.5% for Chat-GPT 4o on the first attempt, 55% on the second, and 47.5% for Claude 3.5. Intra-observer agreement for Chat-GPT 4o across both evaluations was statistically insignificant (p=0.066). In terms of breast parenchymal types, accuracy was 30.19% and 22.64% for the first and second evaluations of Chat-GPT 4o, respectively, and 24.53% for Claude 3.5.

Discussion: There is limited literature on the evaluation of radiographic images by chatbots. However, in one study, mammography images were evaluated using Chat-GPT 4 and Chat-GPT 4o, with an accuracy rate of 66.2%. The same study found a lower accuracy rate for ultrasound examinations at 55.6% (4). In contrast to these studies, which report relatively high accuracy, other studies indicate significantly lower accuracy rates. In a study involving multiple-choice questions with images, an accuracy rate of only 8% was found (5). This study supports such literature by showing that chatbots' visual evaluation abilities fall short of diagnostic accuracy compared to text-based accuracy rates. Additionally, the same mammography exams evaluated by Chat-GPT at different times revealed a lack of intra-observer agreement, indicating high variability and randomness in chatbot image interpretation.

Keywords: Large Language Model (LLM), mammography, artificial intelligence, radiology.