European Journal of Therapeutics

pISSN: 2564-7784
eISSN: 2564-7040

Eur J Ther. 2025;31(1):28-34.
https://doi.org/10.58600/eurjther2599

Original Research

Artificial Intelligence-Based Chatbots' Ability to Interpret Mammography Images: A Comparison of Chat-GPT 4o and Claude 3.5

Betül Nalan Karahan[1], Emre Emekli[1], Mahmut Altuğ Altın[1]

[1] Department of Radiology, Eskişehir Osmangazi University, Faculty of Medicine, Eskişehir, Türkiye

Received: 2025-01-20 **Accepted:** 2025-02-22 **Published Online:** 2025-02-28

Corresponding Author

Betül Nalan Karahan, MD

Address: Eskişehir Osmangazi University,
Faculty of Medicine,Department of
Radiology, Meşelik Yerleşkesi, 26480,
Eskişehir, Türkiye

E-mail: nalanksgl@gmail.com

ABSTRACT

Objectives: The aim of this study is to compare the ability of artificial intelligence-based chatbots, ChatGPT-4o and Claude 3.5, to interpret mammography images. The study focuses on evaluating their accuracy and consistency in BI-RADS classification and breast parenchymal type assessment. It also aims to explore the potential of these technologies to reduce radiologists' workload and identify their limitations in medical image analysis.

Methods: A total of 53 mammography images obtained between January and July 2024 were analyzed, focusing on BI-RADS classification and breast parenchymal type assessment. The same anonymized mammography images were provided to both chatbots under identical prompts.

Results: The results showed accuracy rates for BI-RADS classification ranging from 18.87% to 26.42% for ChatGPT-4o and 18.7% for Claude 3.5. When BI-RADS categories were grouped into benign group(BI-RADS 1,2) and malignant group(BI-RADS 4,5), the combined accuracy was 57.5% for ChatGPT-4o (initial evaluation) and 55% (second evaluation), compared to 47.5% for Claude 3.5. Breast parenchymal type accuracy rates were 30.19% and 22.64% for ChatGPT-4o, and 26.42% for Claude 3.5.

Conclusions: The findings indicate that chatbots demonstrate limited accuracy and reliability in interpreting mammography images. These results highlight the need for further optimization, larger datasets, and advanced training processes to improve their performance in medical image analysis.

Keywords: artificial intelligence, chatbots, ChatGPT-4o, Claude 3.5, mammography, BI-RADS classification, breast parenchymal type, radiology

INTRODUCTION

Artificial intelligence (AI)-based chatbots are widely used today, and their benefits have been studied across various domains, such as writing scientific articles, conducting literature reviews, radiological reporting, and solving radiological cases. In the medical field, they have been evaluated for answering patient inquiries, performing on medical exams, generating medical questions, and more [1-3]. With recent updates, these chatbots can analyze and interpret images. While the exact mechanisms of how chatbots interpret images are not fully understood, it is believed to involve multimodal learning methods and the integration of machine learning algorithms within chatbot frameworks [4-6]. Chatbots can effectively evaluate non-medical images, but interpreting medical and radiological images is more sensitive and requires meticulous testing for model development. Comprehensive assessments of chatbots' performance in analyzing radiological images remain scarce in the literature.

Breast cancer is the most common cancer among women worldwide and the leading cause of death among women aged 25–59 years [7]. Studies have shown that early diagnosis and treatment of breast cancer significantly improve survival rates [8-10]. Mammography is considered the gold standard for breast cancer screening. In cases with diagnostic indications, mammography can be performed regardless of age or age group. The American College of Radiology (ACR) recommends annual mammography screenings for women over 40 years old. ACR's "Breast Imaging Reporting and Data System (BI-RADS)" is a widely accepted system used to describe breast lesions and categorize them into risk groups. The BI-RADS classification enables radiologists to communicate results to referring physicians in a clear and consistent manner, providing final assessments and specific management recommendations [11]. Routine screening mammography or mammography performed based on specific indicationsnfor women over 40 can be time-consuming and labor-intensive for radiologists. Chatbots' ability to interpret mammography examinations could help reduce radiologists' workload.

The aim of this study is to compare the ability of Chat-GPT 4o and Claude 3.5 to interpret mammography images based on BI-RADS classification and breast parenchymal type.

MATERIALS AND METHODS

The study commenced after obtaining approval from the Eskişehir Osmangazi University non-Interventional Clinical Research Ethics Committee (Date 22.10.2024/No: 21). Screening mammography images of women over 40 years old, taken at our hospital between January 1, 2024 and July 1, 2024, were reviewed. Ten patients were planned for each BI-RADS category. However, due to the limited number of reports in BI-RADS category 3, only three patients could be included during this period. A total of 53 mammography images, reported by two radiologists via consensus, were included in the study. Each mammography image consisted of one craniocaudal (CC) and one mediolateral oblique (MLO) standard view.

All images were anonymized and labeled as "mammography." Each mammography (with two views) was uploaded for analysis to both chatbots in separate sessions, and the same prompt was used for both Chat-GPT 4o and Claude 3.5. Chatbots were tasked with assessing the BI-RADS classification and breast parenchymal type:

BI-RADS Classification:

- BI-RADS 0: Requires additional radiological evaluation.
- BI-RADS I: Mammography within normal limits.low
- BI-RADS II: Benign radiological abnormalities.
- BI-RADS III: Low suspicion abnormalities that require follow-up.
- BI-RADS IV: Abnormalities suspicious for malignancy, requiring close follow-up.
- BI-RADS V: High probability of malignancy.

Breast Parenchymal Types:

- Type I: Breast parenchyma predominantly composed of fatty tissue.
- Type II: Scattered fibroglandular densities within fatty breast parenchyma.
- Type III: Heterogeneously dense breast; reduced mammography sensitivity.
- Type IV: Extremely dense breast; lesions may be missed on mammography.

The same process was repeated for Chat-GPT 4o one day apart. While two separate evaluations were conducted for ChatGPT-4o, Claude 3.5 was evaluated only once for the sake of methodological consistency and process simplification. Accuracy rates for BI-RADS classification and breast parenchymal types were calculated for both chatbots. For BI-RADS classification, BI-RADS 1 and 2 were grouped as benign, and BI-RADS 4 and 5 as malignant, as these do not alter patient management.

Observer consistency (intra-observer agreement) for Chat-GPT 4o was also evaluated.

RESULTS

The initial accuracy for BI-RADS classification was calculated as 18.87% for Chat-GPT 4o, increasing to 26.42% on the second evaluation. For Claude 3.5, accuracy was 18.7%. When BI-RADS 1 and 2 were grouped as benign, and BI-RADS 4 and 5 as malignant, the combined accuracy for 40 patients was 57.5% for Chat-GPT 4o initially and 55% on the second evaluation. For Claude 3.5, the accuracy was 47.5% (Table 1).

Intra-observer agreement (ICC) for Chat-GPT 4o was statistically insignificant (p=0.066). Regarding breast parenchymal types, accuracy rates for Chat-GPT 4o were 30.19% and 22.64% for the first and second evaluations, respectively, while for Claude 3.5, it was 26.42% (Table 2).

DISCUSSION

Artificial intelligence-based chatbots like ChatGPT-4o are designed for advanced natural language understanding and generation. These models can process and generate human-like texts thanks to comprehensive pre-training [12]. However, a significant limitation of chatbots is their text-based nature. While image generators such as DALL-E have achieved impressive results in creating visual content, integrating such capabilities into text-based chatbots remains challenging [13]. ChatGPT-4o and other chatbots have recently introduced updates that include image uploading functionality, marking significant progress in their ability to analyze images [6,14-16]. In addition to these advancements, radiologists increasingly utilize the power of artificial intelligence in interpreting medical images.

Table 1. Responses of Chatbots in BI-RADS Classification

	ChatGPT 4o (First response)						**ChatGPT 4o (Second response)**						**Claude 3,5**					
	BI-RADS 0	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5	BI-RADS 0	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5	BI-RADS 0	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5
BI-RADS 0			10					3	1	1	5			1	1		2	6
BI-RADS 1		**4**	3	2	1			**6**	2	1	1			**2**	2		4	2
BI-RADS 2		5	**1**	1	3			2	**2**	1	5			1	**2**		4	3
BI-RADS 3		3							2		1			1			1	1
BI-RADS 4		3	2		**5**			2	1	2	**5**			2	2		**3**	3
BI-RADS 5		3	2		5				3	2	4	**1**	1	2	1		3	**3**

Table 2. Responses of Chatbots in Breast Parenchymal Typing

	ChatGPT 4o (First response)				**ChatGPT 4o (Second response)**				**Claude 3,5**			
	Type I	Type II	Type III	Type IV	Type I	Type II	Type III	Type IV	Type I	Type II	Type III	Type IV
Type I	**1**	3	17			3	18			8	13	
Type II		**2**	16		1		17		1	**4**	13	
Type III			**13**			1	**12**			3	**10**	
Type IV			1			1				1		

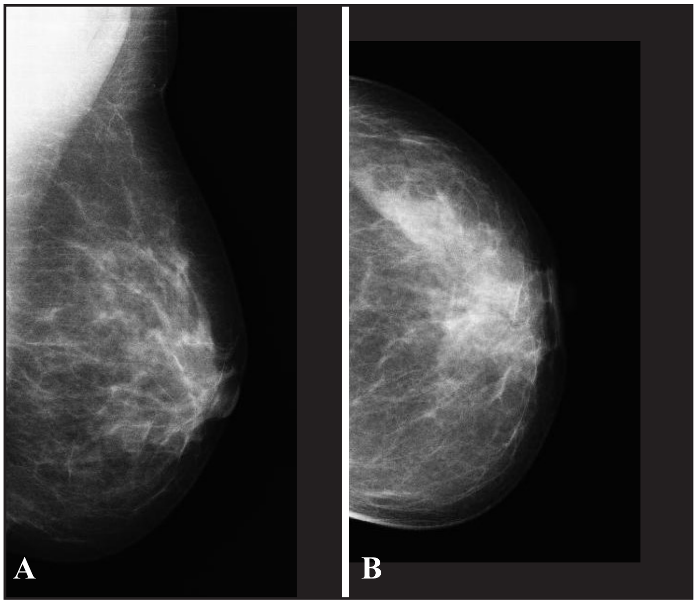

Figure 1. Left MLO (a) and CC (b) mammogram images reported as Type II breast parenchyma and BI-RADS I category by two radiologists. Evaluations by Chat-GPT 4o were as follows first response type III breast parenchyma and BI-RADS IV category, second response type III breast parenchyma and BI-RADS II category. Evaluation by Claude 3.5: Type II breast parenchyma and BI-RADS I category.

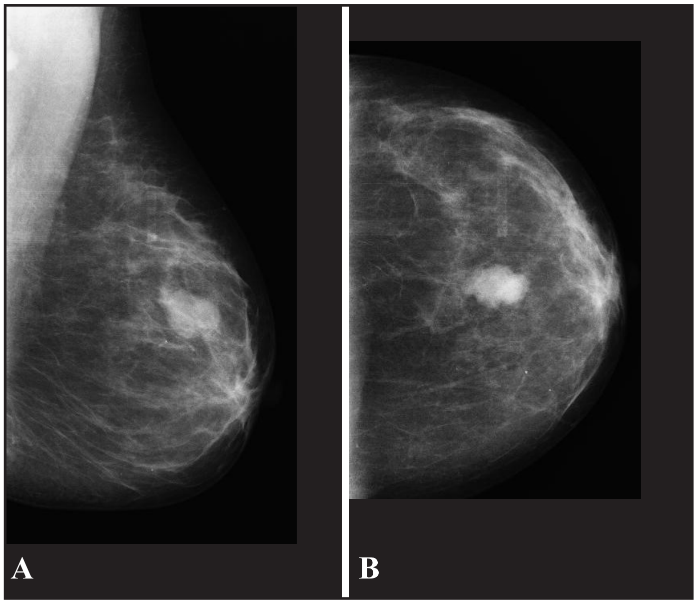

Figure 2. Left MLO (a) and CC (b) mammogram images reported as type II breast parenchyma and BI-RADS V category by two radiologists. Evaluations by Chat-GPT 4o were as follows first response type III breast parenchyma and BI-RADS IV category, second response type III breast parenchyma and BI-RADS IV category. Evaluation by Claude 3.5: Type III breast parenchyma and BI-RADS V category.

This study aims to evaluate routine mammography images, a significant part of radiologists' workload, using chatbots. Chatbots have rapidly advanced in recent years. However, updates related to image generation and interpretation are relatively new. Consequently, there is limited research in the literature on evaluating radiographic images using chatbots.

Studies on image generation indicate that chatbots perform well in generating images from given text inputs [17, 18]. On the other hand, studies on medical image generation suggest that these efforts fall far short of creating realistic images [19, 20]. Moreover, in a study by Shifai et al. [21], GPT-4V was tasked with distinguishing melanoma and benign nevi in dermoscopic images. The overall diagnostic accuracy was found to be 36%. When the researchers considered the three differential diagnoses provided by ChatGPT to be correct, the accuracy increased to 55%. Another study involving ophthalmological images used various imaging modalities. It was noted that ChatGPT failed to identify imaging modalities with high accuracy, with an overall accuracy rate of just 30.5% [22]. These accuracy rates highlight the limitations of chatbots in image interpretation. Similarly, in this study, accuracy rates ranged between 18.87% and 26.42%, aligning with the literature. These findings are significantly lower than the results from convolutional neural network (CNN)-based AI algorithms currently used in the market, highlighting the potential risks of misdiagnosis in clinical use.

In another study involving multiple-choice questions with visual content, the accuracy rate was found to be 8% [23]. This study also supports the literature, showing that chatbots' visual evaluations fall far short of diagnostic accuracy compared to language-based tasks. Furthermore, in this study, text-based ChatGPT-4 demonstrated significantly better diagnostic accuracy than the visual-based GPT-4V, indicating that adding visual data does not directly enhance diagnostic performance [23].

Nguyen et al. [24] conducted a study where mammography and ultrasonography images were evaluated by ChatGPT-4 and ChatGPT-4o. The accuracy rate for mammography images was calculated as 66.2%, while the rate for ultrasonography images was lower at 55.6%, though still relatively high compared to the literature. The higher accuracy rates in this study may be attributed to the use of images from Radiopaedia.org, which includes specific, high-resolution images with defined features for BI-RADS categories. Additionally, chatbots can utilize all metadata associated with the images, contributing to higher accuracy rates.

The results of this study demonstrate that chatbots still have significant limitations in analyzing high-resolution and detail-intensive medical images such as mammography. Considering the diversity and complexity of mammography images used in this study, it can be hypothesized that analyses with high-resolution, specific, and standardized images could improve accuracy. However, such datasets typically require controlled environments and are not widely accessible. Furthermore, the sufficiency and diversity of the datasets used to train models like ChatGPT-4o and Claude 3.5 are crucial factors in accurately interpreting medical images. By incorporating more medical data and visual content during training, accuracy rates could be improved.

Mammography, a critical tool for the early detection of serious conditions like breast cancer, requires precise and accurate results. For chatbots to provide highly accurate interpretations in such critical domains, more extensive training, larger datasets, and model optimizations are needed.

Limitations

This study has several limitations. First, only 53 patients' mammography images were evaluated, and this number could be expanded to include more patients. Due to the insufficient number of patients in the BI-RADS 3 category, only three cases were evaluated, creating an imbalance in statistical analysis. Additionally, this research only utilized ChatGPT-4o and Claude 3.5 chatbots; it could be extended to include other advanced chatbot models. By comparing the performance of different chatbots, the most accurate and reliable model could be identified. The study used radiologists' reports as references, which may also contain errors. To address this, only patients evaluated by consensus between two radiologists were included. Finally, the chatbots conducted evaluations without access to clinical data. However, this approach was chosen to focus solely on the chatbots' capability to analyze images. Since this study involved screening mammography patients, clinical history was not expected to significantly influence reporting.

CONCLUSIONS

Future research could increase the reliability and accuracy of chatbot visual evaluations by using larger datasets. Additionally, the performance of chatbots in interpreting other medical images, such as ultrasonography and MRI, could be explored. Finally, more comprehensive training processes and optimization techniques are required for chatbots to analyze medical images more accurately and reliably.

The accuracy rates obtained in this study indicate that chatbots underperformed in classifying BI-RADS categories and assessing breast parenchymal types in mammography images. These accuracy rates clearly show that chatbots struggle to analyze medical images like mammography accurately. Similarly, the low accuracy rates for breast parenchymal type assessments demonstrate that chatbots fail to recognize such visual details. Additionally, the statistically insignificant intra-observer agreement (ICC) of ChatGPT-4o suggests that chatbots may provide inconsistent results when evaluating the same image at different times, posing a risk to reliability in clinical practice.

The lack of intra-observer consistency underscores the need for further optimization to improve the reliability of chatbots in visual analysis. Such inconsistencies raise concerns about the potential of chatbots in clinical decision-making processes. For AI-based tools to achieve greater reliability and enable chatbots trained on medical images to make human-like, consistent, and accurate decisions, more rigorous training and development will be essential.

Conflict of Interest: The authors declare no conflicts of interest.

Funding: None.

Ethical Approval: Ethical approval was obtained from the Eskişehir Osmangazi University non-Interventional Clinical Research Ethics Committee (Date 22.10.2024/No: 21).

REFERENCES

[1] Tepe M, Emekli E. (2024) Assessing the Responses of Large Language Models (ChatGPT-4, Gemini, and Microsoft Copilot) to Frequently Asked Questions in Breast Imaging: A Study on Readability and Accuracy. Cureus. 9;16(5). https:// doi.org/10.7759/cureus.59960

[2] Fütterer T, Fischer C, Alekseeva A, et al. (2023) ChatGPT in education: global reactions to AI innovations. Sci Rep 15;13(1):15310. https:// doi.org/10.1038/s41598-023-42227-6.

[3] Kıyak YS, Emekli E. (2024) ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J. 18;100(1189):858-865. https:// doi.org/10.1093/postmj/qgae065.

[4] Huang S, Zhang H, Gao Y, Hu Y, Qin Z. (2024) From image to video, what do we need in multimodal LLMs? ArXiv abs/2404.11865 https://doi.org/10.48550/arXiv.2404.11865

[5] Alshehri AS, Lee FL, Wang S. (2023) Multimodal deep learning for scientific imaging interpretation. ArXiv/2309.12460. https://doi.org/10.48550/arXiv.2309.12460

[6] Reizinger P, Ujváry S, Mészáros A, Kerekes A, Brendel W, Huszár F. (2024) Understanding LLMs requires more than statistical generalization. ArXiv preprint arXiv:2405.01964. https://doi.org/10.48550/arXiv.2405.01964

[7] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics. (2021). CA Cancer J Clin. 2021;71:7–33. https://doi.org/10.3322/caac.21654.

[8] Swedish Organised Service Screening Evaluation Group. Reduction in breast cancer mortality from organized service screening with mammography: 1. Further confirmation with extended data. (2006) Cancer Epidemiol Biomarkers Prev. 15(1):45-51. https://doi.org/10.1158/1055-9965.EPI-05-0349.

[9] Kalager M, Haldorsen T, Bretthauer M, Hoff G, Thoresen SO, Adami HO. (2009) Improved breast cancer survival following introduction of an organized mammography screening program among both screened and unscreened women: A population-based cohort study. Breast Cancer Res. 11:R44. https://doi.org/10.1186/bcr2331.

[10] Lauby-Secretan B, Scoccianti C, Loomis D, Benbrahim-Tallaa L, Bouvard V, Bianchini F, et al. (2015) Breast-cancer screening--viewpoint of the IARC Working Group. N Engl J Med. 372:2353–8. https://doi.org/10.1056/NEJMsr1504363.

[11] American College of Radiology. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System Reston, VA: (USA) 2013. Available at: https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/Mammography-Reporting.pdf.

[12] Tepe M, Emekli E. (2024) Decoding medical jargon: The use of AI language models (ChatGPT-4, BARD, microsoft copilot) in radiology reports. Patient Educ Couns. 126:108307. https://doi.org/10.1016/j.pec.2024.108307.

[13] Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL.E 3 for illustrating congenital heart diseases. (2024) J Med Syst. 48(1):54. https://doi.org/10.1007/s10916-024-02072-0.

[14] Adams LC, Busch F, Truhn D, Makowski MR, Aerts HJWL, Bressem KK. What Does DALL-E 2 know about radiology? (2023) J Med Internet Res. 25:e43110. https://doi.org/10.2196/43110.

[15] Bing M. Microsoft Bing Chatbot. (2023)

[16] OpenAI. GPT-4. OpenAI. https://openai.com/index/gpt-4/ Access date: 10.02.2025

[17] Paananen V, Oppenlaender J, Visuri A. Using text-to-image generation for architectural design ideation. (2024) International Journal of Architectural Computing. 22(3):458-474. https://doi.org/10.1177/14780771231222783.

[18] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. (2022) IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042.

[19] Ajmera P, Nischal N, Ariyaratne S, Botchu B, Bhamidipaty KDP, Iyengar KP, Ajmera SR, Jenko N, Botchu R. (2024) Validity of ChatGPT-generated musculoskeletal images. Skeletal Radiol. 53(8):1583-1593. https://doi.org/10.1109/10.1007/s00256-024-04638-y.

[20] Zhu L, Lai Y, Mou W, et al. (2024) ChatGPT's ability to generate realistic experimental images poses a new challenge to academic integrity. J Hematol Oncol. 17:27. https://doi.org/10.1186/s13045-024-01543-8.

[21] Shifai N, van Doorn R, Malvehy J, Sangers TE. (2024) Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol. S0190-9622(24)00076-8. https://doi.org/10.1016/j.

jaad.2023.12.062.

[22] Xu P, Chen X, Zhao Z, Zheng Y, Jin G, Shi D, He M. (2023) Evaluation of a digital ophthalmologist app built by GPT4-V(ision). medRxiv https://doi.org/10.1101/2023.11.27.2329905.

[23] Horiuchi D, Tatekawa H, Oura T, et al. (2024) ChatGPT; diagnostic performance based on textual vs. visual information compared to radiologists; diagnostic performance in musculoskeletal radiology. Eur Radiol. https://doi.org/10.1007/s00330-024-10902-5.

[24] Nguyen D, Rao A, Mazumder A, Succi MD. (2025) Exploring the accuracy of embedded ChatGPT-4 and ChatGPT-4o in generating BI-RADS scores: a pilot study in radiologic clinical practice. 117:110335. https://doi.org/10.1016/j.clinimag.2024.110335.

How to Cite;

Karahan BN, Emekli E, Altın MA (2025) Artificial Intelligence-Based Chatbots' Ability to Interpret Mammography Images: A Comparison of Chat-GPT 4o and Claude 3.5. Eur J Ther. 31(1):28-34. https://doi.org/10.58600/eurjther2599