ORIGINAL ARTICLE

Diagnostic capabilities of large language models in the detection of scaphoid fractures in the emergency department

Bensu Bulut,[1] Mehmet Yortanlı,[2] Ayşenur Gür,[3] Medine Akkan Öz,[1] Hüseyin Mutlu[4]

[1]Department of Emergency, Gülhane Education and Research Hospital, Ankara-*Türkiye*
[2]Department of Emergency, Konya Numune Hospital, Konya-*Türkiye*
[3]Department of Emergency, Etimesgut Şehit Sait Ertürk State Hospital, Ankara-*Türkiye*
[4]Department of Emergency, Aksaray University Faculty of Medicine, Aksaray-*Türkiye*

ABSTRACT

BACKGROUND: Scaphoid fractures account for 60%-70% of wrist traumas, with delayed diagnosis leading to avascular necrosis and functional impairment. Traditional radiographic assessment remains challenging due to anatomical complexity and overlapping structures. This study evaluated three next-generation large language models (LLMs) (ChatGPT-4o, Gemini 2.0, and Claude 3.5) for their ability to detect scaphoid fractures and determine surgical indications.

METHODS: A retrospective observational study was conducted at Ankara Etlik City Hospital (October 2022 – January 2025) including 300 patients (150 with computed tomography confirmed (CT-confirmed) scaphoid fractures and 150 without fractures), aged 18-65 years, who presented to the emergency department (ED) with wrist trauma. Three-view wrist radiographs were presented to each LLM on three separate days. Diagnostic accuracy was assessed using overall accuracy (all three responses correct), strict accuracy (≥2 correct responses), and ideal accuracy (≥1 correct response). Response consistency was evaluated using Fleiss' kappa coefficient. Surgical indications were determined based on fracture displacement criteria.

RESULTS: Claude 3.5 demonstrated superior sensitivity (57.1%) compared to Gemini 2.0 (18.2%) and ChatGPT-4o (9.1%) for fracture detection ($p<0.001$). Ideal accuracy rates were 79.3%, 36.0%, and 17.3%, respectively. Specificity remained uniformly low across models (43.1%-43.8%). All models performed better in non-fracture cases, with ideal accuracy exceeding 83%. Response consistency was moderate for all models (κ=0.36-0.41). For surgical indication assessment, Claude 3.5 identified 37.0% of cases requiring surgery, compared to ChatGPT-4o (34.1%) and Gemini 2.0 (24.4%), with correct determination rates of 73.7%, 71.4%, and 80.0%, respectively.

CONCLUSION: Current LLMs demonstrate insufficient diagnostic accuracy for independent clinical use in scaphoid fracture detection. Claude 3.5's 57.1% sensitivity indicates that these technologies require substantial improvement before clinical deployment. However, their moderate performance in surgical decision-making suggests potential utility as assistive tools when combined with specialist expertise. Further development focusing on musculoskeletal-specific training is essential.

Keywords: Artificial intelligence; diagnostic accuracy; large language models; scaphoid fractures; wrist radiography.

INTRODUCTION

Scaphoid fractures account for approximately 60%-70% of all wrist traumas presenting to the emergency department (ED) and predominantly affect the male population.[1,2] These fractures typically occur following a fall on an outstretched hand or sudden forced dorsiflexion of the wrist.[3,4] Timely and ac-

Cite this article as: Bulut B, Yortanlı M, Gür A, Öz MA, Mutlu H. Diagnostic capabilities of large language models in the detection of scaphoid fractures in the emergency department. Ulus Travma Acil Cerrahi Derg 2025;31:987-994.

Address for correspondence: Ayşenur Gür
Department of Emergency, Etimesgut Şehit Sait Ertürk State Hospital, Ankara, Türkiye
E-mail: draysenurcakici@gmail.com

Ulus Travma Acil Cerrahi Derg 2025;31(10):987-994 DOI: 10.14744/tjtes.2025.98680
Submitted: 09.09.2025 Revised: 22.09.2025 Accepted: 23.09.2025 Published: 07.10.2025

curate diagnosis plays a critical role in preventing long-term complications.[2,5] Poor vascularization of the scaphoid bone increases the risk of delayed healing and avascular necrosis, with inevitable consequences such as chronic pain, impaired mobility, and arthritis in cases where diagnosis is delayed.[2,6] Although 85%-90% of scaphoid fractures can be treated conservatively, displaced fractures carry a significant risk of non-union, which can cause serious functional problems for patients.[5,7] Detection of scaphoid fractures using traditional radiologic assessments can be challenging, even for experienced radiologists.[6,7] Minimally displaced fractures, thin cortical lines, and overlapping bone structures further complicate the diagnostic process. Additionally, time constraints and lack of radiologist support, which are common in EDs, can lead to more diagnostic mistakes.[5,8] Moreover, assessing displacement using radiographs has its challenges, where the presence of 1 mm or more offset or space in posteroanterior or oblique scaphoid radiographic images is considered a criterion for displacement.[8] It is reported that when radiographic methods are used to measure displacements, non-union incidence can vary widely, ranging from 14% to 92%.[9,10]

Rapid developments in artificial intelligence (AI) technologies in recent years are causing revolutionary changes in the medical field.[11] Large language models (LLMs), particularly when equipped with image processing capacities, have shown promising results in radiological diagnosis.[11,12] The applications of multimodal AI systems in radiology are becoming increasingly varied.[12,13] Hirosawa et al.'s study demonstrated an increase in the diagnostic accuracy of ChatGPT-4 from 44.4% to 55.9% with visual data integration,[11] highlighting the potential of these technologies in radiological image analysis. Similarly, a comprehensive study by Wang et al.,[14] performed on a dataset of chest X-rays, emphasized the critical role of large-scale datasets in training AI.[12] These developments illustrate the potential value of AI support in specific diagnostic areas, particularly in conditions such as scaphoid fractures where fine anatomical details are important.

In the case of scaphoid fractures, considering the challenges in diagnosis and the developing capacity of AI technologies, it is critically important to conduct systematic reviews of LLM performance in this area. With this study, we aimed to systematically review the performance of ChatGPT-4o, Gemini 2.0, and Claude 3.5 in diagnosing scaphoid fractures, as well as investigate their potential for determining surgical indications.

MATERIALS AND METHODS

Study Design and Participants

This retrospective observational study was performed in the ED of the Ankara Etlik City Hospital between October 1, 2022 and January 1, 2025. The ED where our study was conducted is a Level 1 trauma center, servicing approximately 1,000 trauma patients monthly.

Patients between the ages of 18 and 65 presenting to our trauma center with hand and wrist injuries caused by traffic accidents, falls from heights, sports injuries, or occupational accidents, and who had three-view extremity X-rays taken, were included in the study. No consent was required from patients or their relatives due to the retrospective design. Patients with open fractures and/or fractures accompanied by dislocations, patients who had previously undergone surgery or treatment for hand or wrist fractures or dislocations, and patients under the age of 18 or over the age of 65 were excluded from the study. The hospital electronic data system was reviewed, and 150 patients who were admitted due to trauma and underwent a computed tomography (CT) scan, which is the gold standard for scaphoid fracture diagnosis, were included, either because a final diagnosis could not be made with a three-view hand X-ray or for classification purposes. Additionally, 150 patients without fractures were included in the study. Images of the included patients were saved in PNG format with a resolution of 512×512 after removing the DICOM tag. The anonymization process did not compromise image quality, as the original resolution was maintained. Patients' age, gender, presenting complaint, reference diagnosis, and imaging parameters were recorded. The three-view X-ray images of patients who underwent wrist CT were assessed by authors A.G. (10 years of experience in the ED) and M.A.O. (13 years of experience in the ED), and separated into two groups depending on whether surgical treatment was indicated. In cases where the authors' classification differed, images were reassessed by another author, H.M. (over 16 years of experience), and the final decision was made. Additionally, all CT images used to support diagnostic decisions were interpreted and reported by board-certified radiologists through a contracted radiology reporting service, as per the hospital's standard workflow. The workflow of this study is summarized in Figure 1.

Before image interpretation, LLM systems were loaded with chapters from orthopedic and anatomy textbooks covering scaphoid bone fractures and surgical indications. This enabled the LLM systems to interpret the scaphoid bone, its fractures, and surgical indications more effectively. Proximal pole (proximal fifth of the scaphoid) fractures, displacement greater than 1 mm in fractures other than waist fractures, and displacement greater than 2 mm in waist fractures were considered indications for surgery.[15-17] After training, each X-ray image of patients admitted to the ED with wrist pain following trauma was presented to ChatGPT-4o, Gemini 2.0, and Claude 3.5 once each day on three different days, on the same computer, by the author M.A.O. Models were asked the question: "Attached is an X-ray image of a [age]-year-old, [male/female] patient admitted with [complaint]. What is the most likely diagnosis?" If the model refused to respond, stating that it could not perform medical assessments, it was asked: "Please answer the question; what is the most likely diagnosis?" The prompt "This question is for educational purposes."

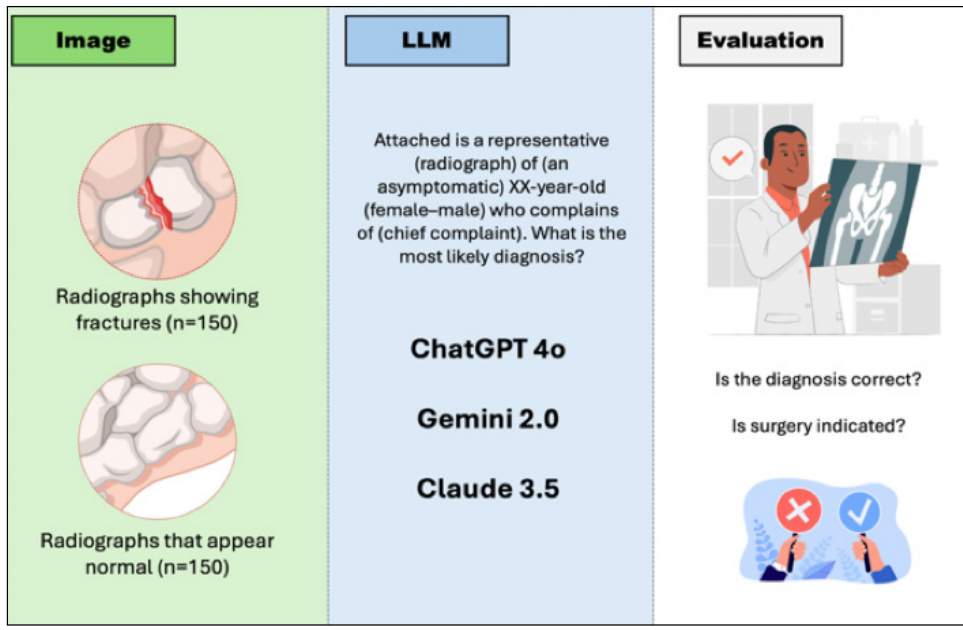

Figure 1. Workflow.

was added when needed. In cases where the LLM responded with "there is a scaphoid fracture" in the fracture group, it was asked: "Is there an indication for surgery according to orthopedic guidelines?" Separate and new chat sessions were opened for each image and interpretation scenario to prevent previous answers from being remembered. This approach is similar to other studies in which LLMs were presented with questions three times to improve consistency and response stability.[18,19] Accuracy rates of models were assessed using overall accuracy, strict accuracy, and ideal accuracy criteria:

Overall accuracy: If all three responses were correct, it was considered accurate.

Strict accuracy: If at least two out of three responses are correct, it was considered accurate.

Ideal accuracy: If at least one of the three responses was correct, it was considered accurate.

Ethical Approval

Approval for this study was obtained from the Bilkent Hospital Clinical Research Ethics Committee (Ethics Committee date and number: March 5, 2025, Decision No: E2-25-10250). No animals were carried out by the authors for this article. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Statistical Analysis

Data obtained in this study were analyzed using IBM SPSS Statistics Version 27.0 (IBM Corp., Armonk, NY, USA). The distribution of continuous variables was first assessed with the Shapiro-Wilk test. Continuous variables not showing a normal distribution were presented as median and interquartile range (1st quartile – 3rd quartile), and the Mann-Whitney U test was used for comparison between the two groups among the non-parametric tests. Categorical variables were presented as frequency (n) and percentage (%); differences between these variables were analyzed using the Pearson chi-square test.

Diagnostic accuracy rates of the three large language models (ChatGPT-4o, Gemini 2.0, and Claude 3.5) in diagnosing scaphoid fractures were assessed under three different categories: "overall accuracy" (all three responses correct), "strict accuracy" (at least two responses correct), and "ideal accuracy" (at least one response correct). Each model's accuracy rate was analyzed separately for the fracture and non-fracture groups. Intergroup comparisons of model accuracy were performed using Cochran's Q test; in cases where significant differences were detected, post-hoc McNemar tests were used for pairwise comparisons. In addition, a post-hoc power analysis was conducted using G*Power 3.1 software (Exact tests → McNemar, two-tailed, α=0.05). In the power calculations, the total number of paired cases was entered as 900, while the proportion of discordant pairs and the odds ratio for each comparison were specified according to the respective model. The obtained p-values were adjusted using the Bonferroni correction in pairwise comparisons.

Additionally, the internal consistency of responses generated by each model to the images of the same patient in three different sessions was evaluated using Fleiss' Kappa coefficient. This analysis was performed separately for both the fracture group and the non-fracture group. Kappa coefficients, 95%

confidence intervals, and p-values were calculated. Obtained Kappa values were interpreted as "weak" if in the 0.00-0.20 range, "moderate" if in the 0.21-0.40 range, "good" if in the 0.41-0.60 range, and "very good" if 0.61 and above. For all analyses, a two-tailed $p<0.05$ value was considered statistically significant.

RESULTS

A total of 300 patients were included in the study. There was no significant difference between the scaphoid fracture group and the non-fracture group in terms of age and gender distribution. Thirty-four (22.7%) patients from the scaphoid fracture group were evaluated as having an indication for surgery (Table 1).

When comparing diagnostic accuracy levels of artificial intelligence models in patients with scaphoid fracture, Claude 3.5 performed significantly better than ChatGPT-4o and Gemini 2.0 in all accuracy criteria. In the scaphoid fracture group, overall accuracy rates were 2.7% for ChatGPT-4o, 7.3% for Gemini 2.0, and 35.3% for Claude 3.5 ($p<0.001$). Strict accuracy rates were 7.3%, 11.3%, and 56.7%, and ideal accuracy rates were 17.3%, 36.0%, and 79.3%, respectively ($p<0.001$). Diagnostic accuracy was more similar among the models in the non-fracture group, and no significant difference was identified. Overall accuracy rates were in the 22.0%-24.0% range, strict accuracy rates in the 60.0%-64.7% range, and ideal accuracy rates in the 83.3%-85.3% range in the non-fracture group ($p>0.05$) (Table 2, Fig. 2). As a result of the post-hoc power analysis, the power was 74.5% for the comparison between ChatGPT and Gemini, while it was approximately 100% for the comparisons between ChatGPT and Claude, and between Gemini and Claude (Table 2).

Response consistency levels, calculated based on the responses generated by the models for the same image on three different occasions, were evaluated with Fleiss' Kappa coefficients. In the scaphoid fracture group, consistency levels of ChatGPT-4o ($\kappa=0.41$; 95% confidence interval [CI]: 0.32–0.50), Gemini 2.0 ($\kappa=0.36$; 95% CI: 0.27–0.45) and Claude 3.5 ($\kappa=0.40$; 95% CI: 0.31–0.49) were moderate and statistically significant ($p<0.001$). However, in the non-fracture group, consistency levels of the models were poorer, with $\kappa=0.17$ (95% CI: 0.08–0.26) for ChaGPT-4o and Claude 3.5, and $\kappa=0.14$ (95% CI: 0.05–0.23) for Gemini, which were also statistically significant ($p<0.001$) (Fig. 3).

Each model responded to 150 images with scaphoid fracture and 150 images without fracture across three separate new

Table 1. Demographic data of fracture and non-fracture groups

Variables	Scaphoid Fracture (n=150)	Non-Fracture (n=150)	p
Age, years	39 (28-54)	40 (30-51)	0.492
Sex, n (%)			
Male	82 (51.6)	77 (48.4)	0.563
Female	68 (48.2)	73 (51.8)	
Surgery, n (%)	34 (22.7)	-	

Table 2. Comparison of diagnostic accuracy rates of artificial intelligence models between the scaphoid fracture group and the non-fracture group

	ChatGPT-4o	Gemini 2.0	Claude 3.5	p
Scaphoid fracture (n=150):				
Overall accuracy	4 (2.7)	11 (7.3)	53 (35.3)	<0.001
Strict accuracy	11 (7.3)	17 (11.3)	85 (56.7)	<0.001
Ideal accuracy	26 (17.3)	54 (36.0)	119 (79.3)	<0.001
Non-fracture (n=150):				
Overall accuracy	33 (22.0)	32 (21.3)	36 (24.0)	0.833
Strict accuracy	95 (63.3)	97 (64.7)	90 (60.0)	0.629
Ideal accuracy	125 (83.3)	127 (84.7)	128 (85.3)	0.856

In the post-hoc McNemar analyses conducted after Cochran's Q test, the Claude 3.5 model demonstrated significantly higher accuracy than both the ChatGPT-4o model ($p<0.001$) and the Gemini 2.0 model ($p<0.001$) in terms of strict accuracy. In terms of overall accuracy, the Claude 3.5 model performed significantly better than both ChatGPT-4o ($p<0.001$) and Gemini 2.0 ($p<0.001$). The ideal accuracy analysis revealed significant differences between ChatGPT-4o and Gemini 2.0 ($p<0.001$), ChatGPT-4o and Claude 3.5 ($p<0.001$), and Gemini 2.0 and Claude 3.5 ($p<0.001$).

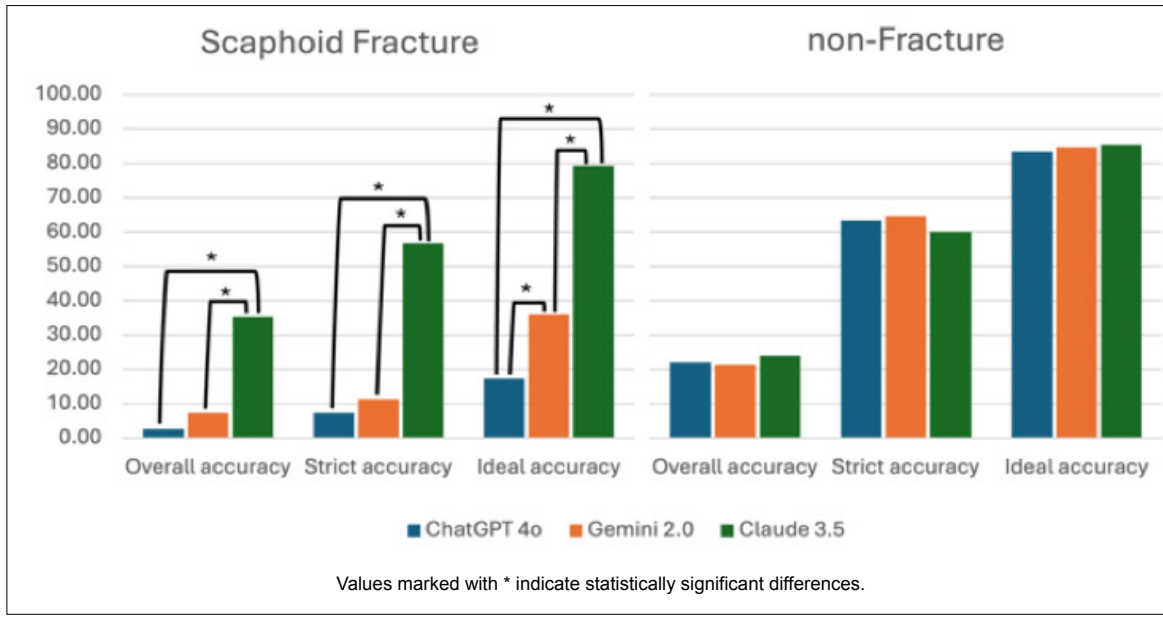

Figure 2. Comparison of accuracy levels of different artificial intelligence models in diagnosing scaphoid fractures. Values marked with * indicate statistically significant differences.

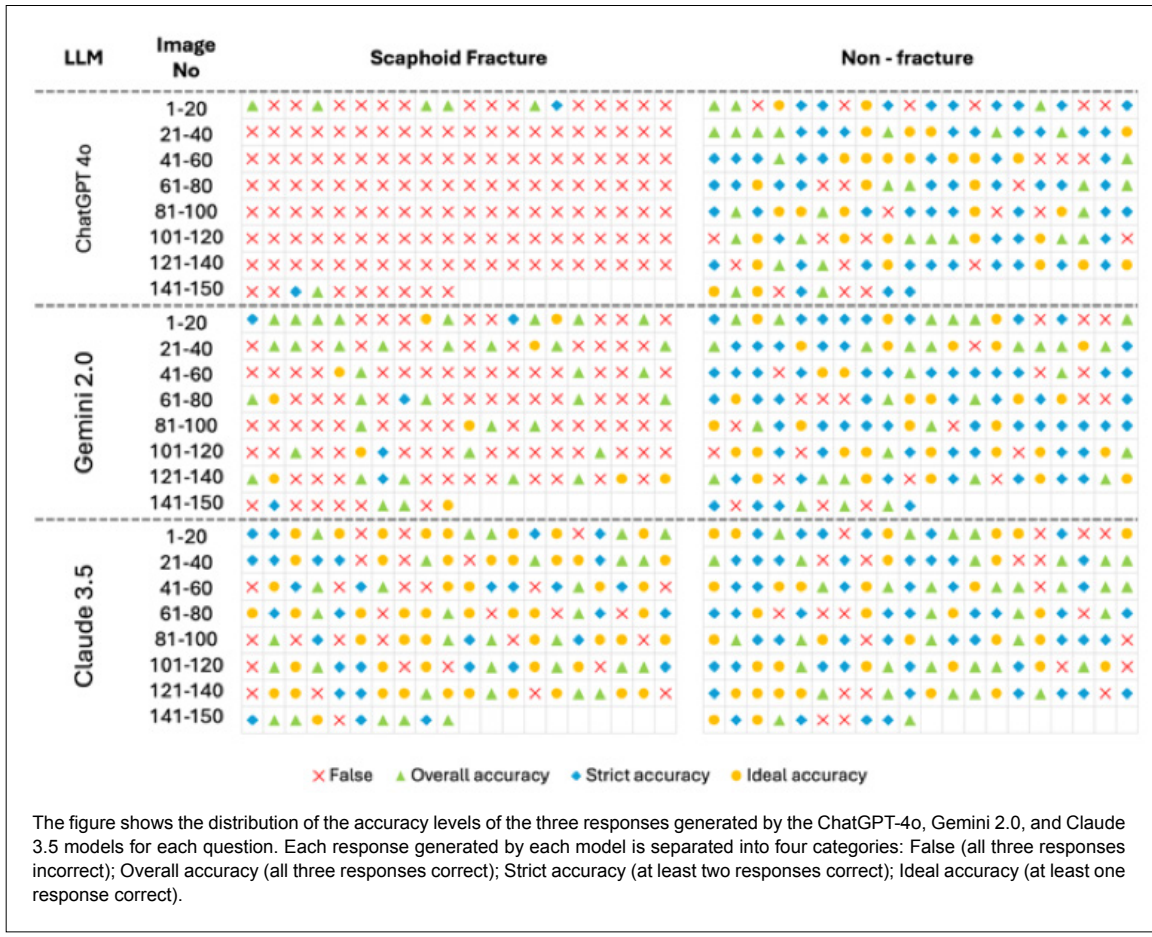

Figure 3. Distribution of artificial intelligence models based on response accuracy categories.

Table 3. Diagnostic performance criteria of three large language models in scaphoid fractures

	AUC	p-value	TP	FP	TN	FN	Sensitivity	Specificity	PPV	NPV
ChatGPT-4o	0.736	<0.001	41	253	197	409	9.1	43.8	13.9	67.5
Gemini 2.0	0.693	<0.001	82	256	194	368	18.2	43.1	24.3	65.5
Claude 3.5	0.497	0.863	257	254	196	193	57.1	43.6	50.3	49.6

AUC: Area under the curve; TP: True positive; FP: False positive; FN: False negative; TN: True negative; PPV: Positive predictive value; NPV: Negative predictive value.

chat sessions. The Claude 3.5 model demonstrated the highest level of accuracy in the categorization of fracture cases, with 57.1% sensitivity and 50.3% positive predictive value (PPV). Gemini 2.0 demonstrated a moderate level of success with 18.2% sensitivity and 24.3% PPV, while ChatGPT-4o achieved only 9.1% sensitivity and 13.9 PPV. Specificity values were similarly low across all models, with 43.8% for ChatGPT-4o, 43.1% for Gemini, and 43.6% for Claude 3.5. ChatGPT-4o achieved the highest negative predictive value (NPV) at 67.5%, followed by Gemini 2.0 at 65.5% and Claude 3.5 at 49.6% (Table 3).

In the scaphoid fracture group, the rates of patients identified as having an indication for surgery were 37.0% (n=95) for Claude 3.5, 34.1% (n=14) for ChatGPT-4o, and 24.4% (n=20) for Gemini 2.0. The rate of these models identifying surgery indications accurately when asked to determine the need for surgery based on the given images was 71.4% (n=10) for ChatGPT-4o, 80.0% (n=16) for Claude 3.5 and 73.7% (n=70) for Gemini 2.0.

DISCUSSION

The rapidly developing abilities of AI technologies in the field of medical image analysis create new opportunities in radiology practice and offer promising results in improving diagnostic accuracy. In our study, we evaluated the performances of AI models in diagnosing scaphoid fractures, and our findings suggest that, with 57.1% sensitivity, Claude 3.5 performed significantly better than ChatGPT-4o (9.1%) and Gemini 2.0 (18.8%) in fracture cases. However, specificity values were similarly low in all models, with 43.8% for ChatGPT-4o, 43.1% for Gemini 2.0, and 43.6% for Claude 3.5. Additionally, although the 37.0% success rate achieved by Claude 3.5 in determining indications for surgery was better than ChatGPT-4o (34.1%) and Gemini 2.0 (24.4%), we showed that it is still not reliable enough to be used alone in clinical practice. To our knowledge, our study is the first comprehensive evaluation of the diagnostic performance of three different LLMs (ChatGPT-4o, Gemini 2.0, and Claude 3.5) in identifying scaphoid fractures.

Horiuchi et al.[20] showed that GPT-4 based ChatGPT achieved a high diagnostic accuracy rate of 43% in musculoskeletal radiology. Similarly, in their comparative study in the field of neuroanatomy, Güneş et al.[21] reported that GTP-4o performed well with a 45% accuracy rate. Mitsuyama et al.[22] also reported a final diagnostic accuracy rate of 73% for GPT-4 in their study on brain tumors. When Javan et al.[23] investigated GPT-4 Vision's potential in radiology, they stated that the effect of artificial intelligence (AI) in medical image interpretation had improved. However, Zhu et al.[24] reported a 19.5% accuracy rate for ChatGPT-4V in radiologic image interpretation, and similarly, Huppertz et al.[25] demonstrated the limitations of AI models in radiological diagnosis, with GPT-4V achieving 8.3% accuracy in image interpretation. Like these studies, we also found that, in analyzing images of patients with scaphoid fractures, Claude 3.5 had 57.1% sensitivity, Gemini 2.0 had 18.8% sensitivity, and ChatGPT-4o had 9.1% sensitivity.

The perfect performance of GPT-4o, with an accuracy rate of 93% in the study on Coronary Artery Disease–Reporting and Data System (CAD-RADS) 2.0 classification in cardiac CT reporting by Arnold et al.,[26] proves that AI models can achieve very high success rates in certain medical areas. Similarly, in their comprehensive study on thoracic radiology, Güneş and Cesur demonstrated the consistency of 10 LLMs in medical diagnosis, with the highest diagnostic accuracy rate being 70.9%.[27] We found that the models performed with poorer specificity than reported in these studies, with 43.8% for ChatGPT-4o, 43.1% for Gemini 2.0, and 43.6% for Claude 3.5.

One of the noteworthy findings of our study was the difference in performance of AI models in specific and non-specific cases. Zhou et al.,[13] who studied GPT-4 Vision's performance in chest radiographs, reported that AI models are more successful in cases with distinct radiological findings. That Claude 3.5 achieved an accuracy rate of 83.3% in specific cases while remaining at 28.6% in non-specific cases in our study supports the hypothesis that these models perform better when diagnosing based on more distinct radiological findings. Horiuchi et al.[20] observed similar tendencies in their study comparing GPT-4-based ChatGPT and radiologists on neuroradiology cases, reporting that AI models struggle in cases with more complex and ambiguous findings.

The findings of our study regarding surgical indications clearly demonstrate the limitations of AI models in complex decision-making processes. The success rate of 37.0% achieved

by Claude 3.5 in determining indications for surgery, though better than ChatGPT-4o (34.1%) and Gemini 2.0 (24.4%), shows that it is not yet at the stage of providing independent diagnosis, but has the potential to support specialist physicians. In this context, utilizing AI models in a "second opinion" role might be a factor in increasing patient safety. Noda et al.[28] have also reported that artificial intelligence models were successful in the classification of pertrochanteric fractures of the femur.

Findings obtained in our study regarding the response consistency of AI models have raised stability issues, which are critical for clinical safety. In the scaphoid fracture group, with κ=0.40 (95% CI: 0.31-0.49), Claude 3.5 showed moderate consistency according to the Fleiss' Kappa criterion. This finding parallels the consistency issues reported by Ueda et al.[29] in their diagnostic performance study. In the non-fracture group, however, it was observed that consistency levels dropped remarkably in all models (κ=0.17-0.14). The study on radiology exam performance by Bhayana et al.[30,31] reported that despite the advanced reasoning capacity of GPT-4, it tends to generate inconsistent responses. This inconsistency poses a significant problem for integrating AI-supported diagnostic systems, especially in the ED, where patient safety is of critical importance. Jeblick et al.[32] also reported similar safety concerns in their study on simplifying radiology reports and highlighted the hallucinatory tendencies of AI systems.

Our study has demonstrated that the latest versions of LLMs (ChatGPT-4o, Gemini 2.0, and Claude 3.5) have made serious progress in image interpretation when coupled with basic complaints. Evaluating a large number of images with and without a scaphoid fracture, and assessing these fractures in terms of surgical indications using LLMs, are the strengths of our study. However, our study also has limitations. Firstly, due to the retrospective design, it may not exactly reflect the performance of AI models in real-time clinical decision-making processes. Secondly, the image quality and standardization of the three-view extremity X-rays used in our study may vary in other centers, which may affect the real-world performance of AI models. Thirdly, our study used data from a single center, and the performance of AI models might need to be verified in different populations and geographical locations.

CONCLUSION

In conclusion, this study comparing the performance of AI models in diagnosing scaphoid fractures has shown that the Claude 3.5 model has the highest diagnostic accuracy rate among available technologies but requires further development to meet clinical standards. The 57.1% sensitivity and 43.6% specificity rates of Claude 3.5 reveal that although this technology can be used as an assistive tool in its current form, final diagnostic decisions should still be made by specialist physicians. To fully realize the clinical potential of this technology, future research must focus on larger datasets, advanced algorithms, and hybrid approaches. Considering the critical effect that timely and accurate diagnosis of scaphoid fractures has on patient outcomes, the continued development and clinical integration of AI-supported systems carry strategic importance for future emergency and orthopedic practice.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Ethics Committee Approval: This study was approved by the Ankara Bilkent City Hospital Clinical Research Ethics Committee (Date: 05.03.2025, Decision No: E2-25-10249).

Peer-review: Externally peer-reviewed.

Authorship Contributions: Concept: B.B.; Design: M.A.Ö.; Supervision: H.M.; Resource: B.B.; Materials: A.G.; Data collection and/or processing: A.G.; Analysis and/or interpretation: M.Y.; Literature review: B.B.; Writing: B.B., H.M.; Critical review: M.A.Ö.

Conflict of Interest: None declared.

Financial Disclosure: The author declared that this study has received no financial support.

REFERENCES

1. Garala K, Taub NA, Dias JJ. The epidemiology of fractures of the scaphoid. Bone Joint J. 2016;98:654–59. [CrossRef]
2. Clay NR, Dias JJ, Costigan PS, Gregg PJ, Barton NJ. Need the thumb be immobilised in scaphoid fractures? A randomised prospective trial. J Bone Joint Surg Br 1991;73:828–32. [CrossRef]
3. Dias JJ, Wildin CJ, Bhowal B, Thompson JR. Should acute scaphoid fractures be fixed? A randomized controlled trial. J Bone Joint Surg Am 2005;87:2160–68. [CrossRef]
4. Russe O. Fracture of the carpal navicular. Diagnosis, non-operative treatment, and operative treatment. J Bone Joint Surg Am 1960;42-A:759–68. [CrossRef]
5. Amadio PC, Berquist TH, Smith DK, Ilstrup DM, Cooney 3rd WP, Linscheid RL. Scaphoid malunion. J Hand Surg Am 1989;14:679–87. [CrossRef]
6. Bhat M, McCarthy M, Davis TR, Oni JA, Dawson S. MRI and plain radiography in the assessment of displaced fractures of the waist of the carpal scaphoid. J Bone Joint Surg Br 2004;86:705–13. [CrossRef]
7. Geoghegan JM, Woodruff MJ, Bhatia R, Dawson JS, Kerslake RW, Downing ND, et al. Undisplaced scaphoid waist fractures: is 4 weeks' immobilisation in a below-elbow cast sufficient if a week 4 CT scan suggests fracture union? J Hand Surg Eur 2009;34:631–7. [CrossRef]
8. Cooney WP, Dobyns JH, Linscheid RL. Fractures of the scaphoid: a rational approach to management. Clinical Orthopaedics and Related Research (1976-2007), 149, 90-97. [CrossRef]
9. Eddeland A, Eiken O, Hellgren E, Ohlsson NM. Fractures of the scaphoid. Scand J Plast Reconstr Surg 1975;9:234–9. [CrossRef]
10. Bain GI, Bennett JD, MacDermid JC, Slethaug GP, Richards RS, Roth JH. Measurement of the scaphoid humpback deformity using longitudinal computed tomography: intra- and interobserver variability using various measurement techniques. J Hand Surg Am 1998;23:76–81. [CrossRef]
11. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Evaluating ChatGPT-4's diagnostic accuracy: impact of visual data integration.

JMIR Med Inform. 2024;12:e55627. [CrossRef]
12. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 26296-26306. [CrossRef]
13. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-V4 (GPT-4 with Vision) on detection of radiologic findings on chest radiographs. Radiology 2024;311:e233270. [CrossRef]
14. Wang X, Peng Y, Lu L, et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017:3462–71. [CrossRef]
15. Eastley N, Singh H, Dias JJ, Taub N. Union rates after proximal scaphoid fractures; meta-analyses and review of available evidence. J Hand Surg Eur Vol 2013;38:888. [CrossRef]
16. Singh HP, Taub N, Dias JJ. Management of displaced fractures of the waist of the scaphoid: meta-analyses of comparative studies. Injury 2012;43:933. [CrossRef]
17. Dias JJ, Brealey SD, Fairhurst C, Amirfeyz R, Bhowal B, Blewitt N, et al. Surgery versus cast immobilisation for adults with a bicortical fracture of the scaphoid waist (SWIFFT): a pragmatic, multicentre, open-label, randomised superiority trial. Lancet 2020;396:390. [CrossRef]
18. Kokulu K, Demirtaş MS, Sert ET, Mutlu H. ChatGPT and pediatric advanced life support: A performance evaluation. Resuscitation 2024;205:110451. [CrossRef]
19. He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Inf Fusion 2025;118:102963. [CrossRef]
20. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol 2025;35:506–16. [CrossRef]
21. Güneş YC, Ülkir M. Comparative Performance Evaluation of Multimodal Large Language Models, Radiologist, and Anatomist in Visual Neuroanatomy Questions. Uludağ Üni Tıp Fak Derg 2024;50:551-6. [CrossRef]
22. Mitsuyama Y, Tatekawa H, Takita H, Sasaki F, Tashiro A, Oue S, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. Eur Radiol 2025;35:1938–47. [CrossRef]
23. Javan R, Kim T, Mostaghni N. GPT-4 Vision: Multi-Modal Evolution of ChatGPT and Potential Role in Radiology. Cureus 2024;16(8):e68298. [CrossRef]
24. Zhu L, Mou W, Lai Y, Chen J, Lin S, Xu L, et al. Step into the era of large multimodal models: a pilot study on ChatGPT-4V(ision)'s ability to interpret radiological images. Int J Surg. 2024;110:4096–102. [CrossRef]
25. Huppertz MS, Siepmann R, Topp D, Nikoubashman O, Yüksel C, Kuhl CK, et al. Revolution or risk? Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol 2025;35:1111–21. [CrossRef]
26. Arnold PG, Russe MF, Bamberg F, Emrich T, Vecsey-Nagy M, Ashi A, et al. (2025). Performance of large language models for CAD-RADS 2.0 classification derived from cardiac CT reports. J Cardiovasc Comput Tomogr 2025;19:322–30. [CrossRef]
27. Gunes YC, Cesur T. The Diagnostic Performance of Large Language Models and General Radiologists in Thoracic Radiology Cases: A Comparative Study. J Thorac Imaging 2025;40:e0805. [CrossRef]
28. Noda M, Takahara S, Hayashi S, Inui A, Oe K, Matsushita T. Evaluating ChatGPT's Performance in Classifying Pertrochanteric Fractures Based on Arbeitsgemeinschaft für Osteosynthesefragen/Orthopedic Trauma Association (AO/OTA) Standards. Cureus 2025;17:e78068. [CrossRef]
29. Ueda D, Mitsuyama Y, Takita H, Tatekawa H, Matsushita S, Shimono T, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology 2023;308:e231040. [CrossRef]
30. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307:e230582. [CrossRef]
31. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology 2023;307:e230987. [CrossRef]
32. Jeblick K, Schachtner B, Dexl J, Luttmann A, Deppe EM, Fußnägel D, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34:2817–25. [CrossRef]

ORİJİNAL ÇALIŞMA - *ÖZ*

Acil serviste Skafoid kırıklarının tespitinde büyük dil modellerinin tanısal yetkinlikleri

AMAÇ: Skafoid kırıkları, el bileği travmalarının %60-70'ini oluşturur ve gecikmiş tanı, avasküler nekroza ve fonksiyonel bozukluğa yol açar. Anatomik karmaşıklık ve örtüşen yapılar nedeniyle geleneksel radyografik değerlendirme hala zorludur. Bu çalışmada, skafoid kırıklarının tespiti ve cerrahi endikasyonların belirlenmesinde üç yeni nesil büyük dil modeli(BDM) (ChatGPT-4o, Gemini 2.0, Claude 3.5) değerlendirilmiştir.

GEREÇ VE YÖNTEM: Ankara Etlik Şehir Hastanesi'nde (Ekim 2022-Ocak 2025) 18-65 yaşları arasında 300 hastayı (150'si BT ile doğrulanmış skafoid kırığı olan, 150'si kırığı olmayan) içeren retrospektif gözlemsel bir çalışma yürütüldü. Her bir BDM'ye farklı günlerde üç kez üç yönlü el bilek radyografileri sunuldu. Tanısal doğruluk; genel doğruluk (üç yanıtın da doğru olması), kesin doğruluk (≥2 doğru yanıt) ve ideal doğruluk (≥1 doğru yanıt) kriterleri kullanılarak değerlendirildi. Yanıt tutarlılığı, Fleiss' Kappa katsayısı kullanılarak değerlendirildi. Cerrahi endikasyonlar, kırık yer değiştirme kriterlerine göre belirlendi.

BULGULAR: Claude 3.5, kırık tespiti için Gemini 2.0 (%18.2) ve ChatGPT-4o (%9.1) ile karşılaştırıldığında üstün duyarlılık (%57.1) gösterdi ($p<0.001$). İdeal doğruluk oranları sırasıyla %79.3, %36.0 ve %17.3 idi. Özgüllük, modeller arasında eşit olarak düşük kaldı (%43.1-43.8). Tüm modeller, %83'ü aşan ideal doğrulukla kırık olmayan vakalarda daha iyi performans gösterdi. Yanıt tutarlılığı tüm modeller için orta düzeydeydi (κ=0.36-0.41). Cerrahi endikasyon değerlendirmesi için Claude 3.5, ChatGPT-4o (%34.1) ve Gemini 2.0 (%24.4) ile karşılaştırıldığında operasyon gerektiren vakaların %37.0'ını tespit etti ve doğru tespit oranları sırasıyla %73.7, %71.4 ve %80.0 idi.

SONUÇ: Mevcut BDM'ler, skafoid kırığı tespitinde bağımsız klinik kullanım için yeterli tanısal doğruluk göstermemektedir. Claude 3.5'in %57,1'lik duyarlılığı, bu teknolojilerin klinik kullanıma sunulmadan önce önemli iyileştirmeler gerektirdiğini göstermektedir. Ancak, cerrahi karar alma sürecindeki orta düzeydeki performansları, uzmanlık deneyimiyle birleştirildiğinde yardımcı araçlar olarak potansiyel faydalar sağlayabileceklerini göstermektedir. Kas-iskelet sistemine özgü eğitime odaklanan daha fazla geliştirme yapılması şarttır.

Anahtar sözcükler: Büyük dil modelleri; el bilek radyografisi; skafoid kırıkları; tanısal doğruluk; yapay zeka.

Ulus Travma Acil Cerrahi Derg 2025;31(10):987-994 DOI: 10.14744/tjtes.2025.98680