

DIAGNOSTIC EVALUATION OF COMPUTED
TOMOGRAPHY IMAGES IN PATIENTS WITH SUSPECTED
LUNG LESIONS USING GPT-4O: A COMPARISON OF
TEMPERATURE SETTINGS AND DIAGNOSTIC
ACCURACY

Omer Yusuf Erdurmus^{1✉}, Eser Uyanik², Aykut Guruf³

Received on June 6, 2025

Presented by D. Damianov, Member of BAS, on July 29, 2025

Abstract

Lung cancer remains one of the leading causes of cancer-related deaths worldwide. While computed tomography (CT) imaging plays a crucial role in the detection of lung lesions, the potential use of artificial intelligence (AI) tools in assisting diagnosis is increasingly being explored. This study evaluates the diagnostic performance of GPT-4o, a multimodal AI model, in detecting lung lesions on non-contrast chest CT images of patients admitted to the emergency department.

We retrospectively analyzed CT scans of 112 adult patients (56 with radiologically confirmed lung lesions and 56 healthy controls) admitted to the emergency department in 2024. GPT-4o was prompted to interpret four selected CT slices per patient at three different temperature settings (low, medium, high). Diagnoses were compared to radiologist assessments for accuracy.

GPT-4o demonstrated varying diagnostic performance depending on temperature settings. Sensitivity and specificity at medium temperature were 62.5% and 64.29%, respectively. The highest accuracy (75%) was achieved at low temperature, while high temperature resulted in reduced performance (accuracy: 43.75%). Misinterpretation of lesion location was more common at higher temperatures. GPT-4o demonstrates moderate diagnostic performance in evaluating non-contrast chest CT scans for lung lesions. Its accuracy is

<https://doi.org/10.7546/CRABS.2025.09.11>

highest when using low temperature settings, which favour more conservative and consistent image interpretations. However, the tool showed limitations in detecting lesion locations and in overall sensitivity and specificity.

GPT-4o shows potential as a supportive tool in the detection of lung lesions on CT scans, particularly at low temperature settings. However, its diagnostic performance is moderate and should not replace expert radiologist evaluation. Future improvements and prospective validation studies are needed before routine clinical use.

Key words: artificial intelligence, computed tomography, ChatGPT, lung lesion, diagnostic accuracy

Introduction. Lung cancer is the most common malignancy both in the world and in Türkiye [1,2]. Early diagnosis and treatment are very important in these lung lesion and lung imaging techniques, which are constantly improving in the light of technological developments. Today, artificial intelligence programs using deep learning algorithms are used to accelerate clinical diagnosis processes and increase accuracy, and are used in the literature not only in lung imaging but also in other imaging methods [3–5]. ChatGPT (OpenAI, L.L.C.) is an advanced web-based chatbot using artificial intelligence and a wide language model, and the number of studies in the literature using this artificial intelligence program in imaging evaluation is increasing day by day [3,6–9]. However, the routine use of ChatGPT in clinical practice and the advantages and disadvantages of artificial intelligence technologies are still being discussed today [10]. In our study, tomography sections of patients who applied to the emergency department, underwent chest CT, and were diagnosed with lung lesion as a result of their evaluation by radiologists were evaluated with GPT-4o for diagnostic purposes.

Material and methods. This study retrospectively reviewed patients from Ordu State Hospital Emergency Department (January-December 2024) who underwent chest CT and were diagnosed with a lung lesion. Sample size calculation was conducted with 95% statistical power and type I error probability of $\alpha = 0.05$. To detect an 80% success rate for the tested diagnostic method compared to the 100% success rate of the reference group, it was determined that a total of 112 cases, with a minimum of 56 cases per group, should be included in the study. A total of 112 cases, equally divided between lung lesion and control groups, were analyzed. Four CT slices per case were evaluated by GPT-4o, which provided two likely diagnoses. Radiologists assessed the accuracy of GPT-4o's reports, ensuring standardized case selection and analysis (Fig. 1). Inclusion criteria for the study were determined as patients aged 18 years and over, patients who underwent chest CT imaging and detected a lung lesion. Exclusion criteria were chest CT with contrast, presence of foreign body, presence of pleural effusion, presence of artifactual areas, and lesion size of 1 cm or less.

GPT-4o evaluates visual changes based on temperature settings, ranging from conservative comments at low temperatures (0.0–0.4) to balanced analysis

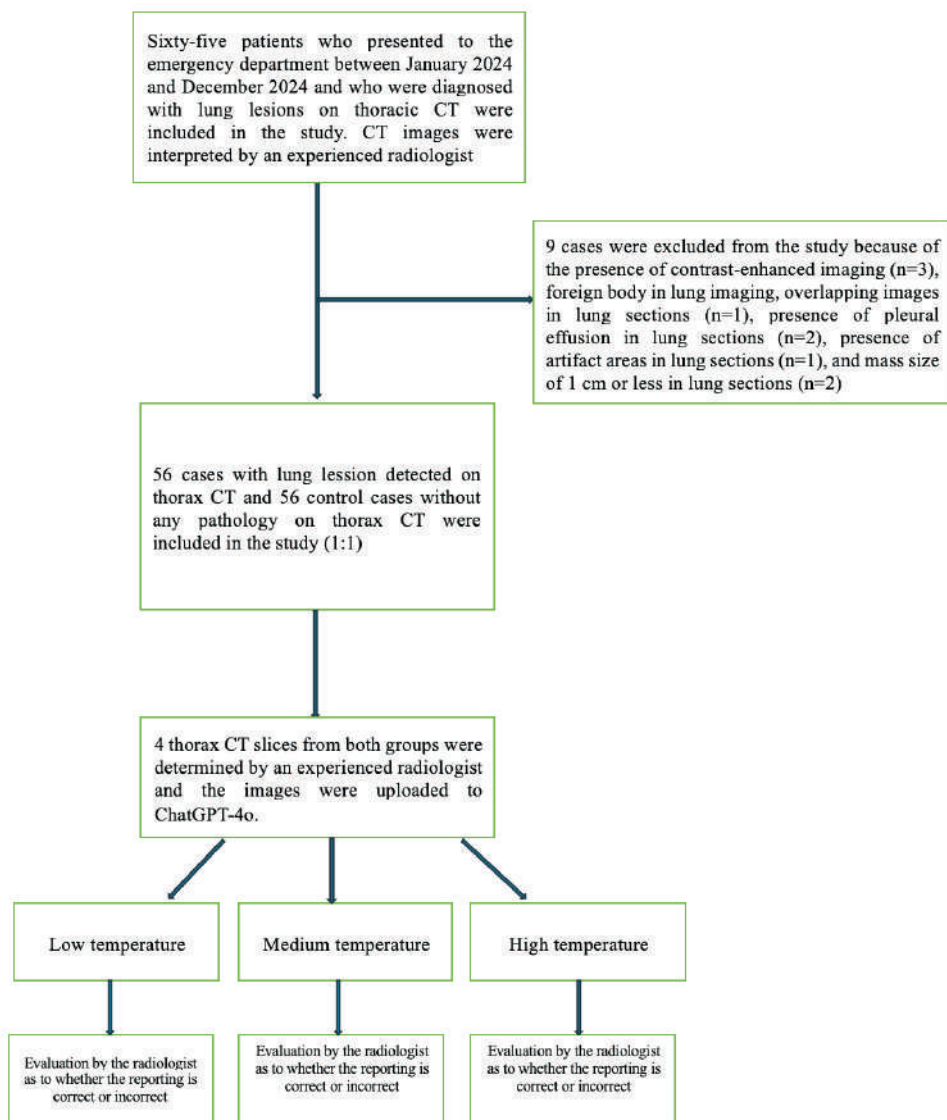
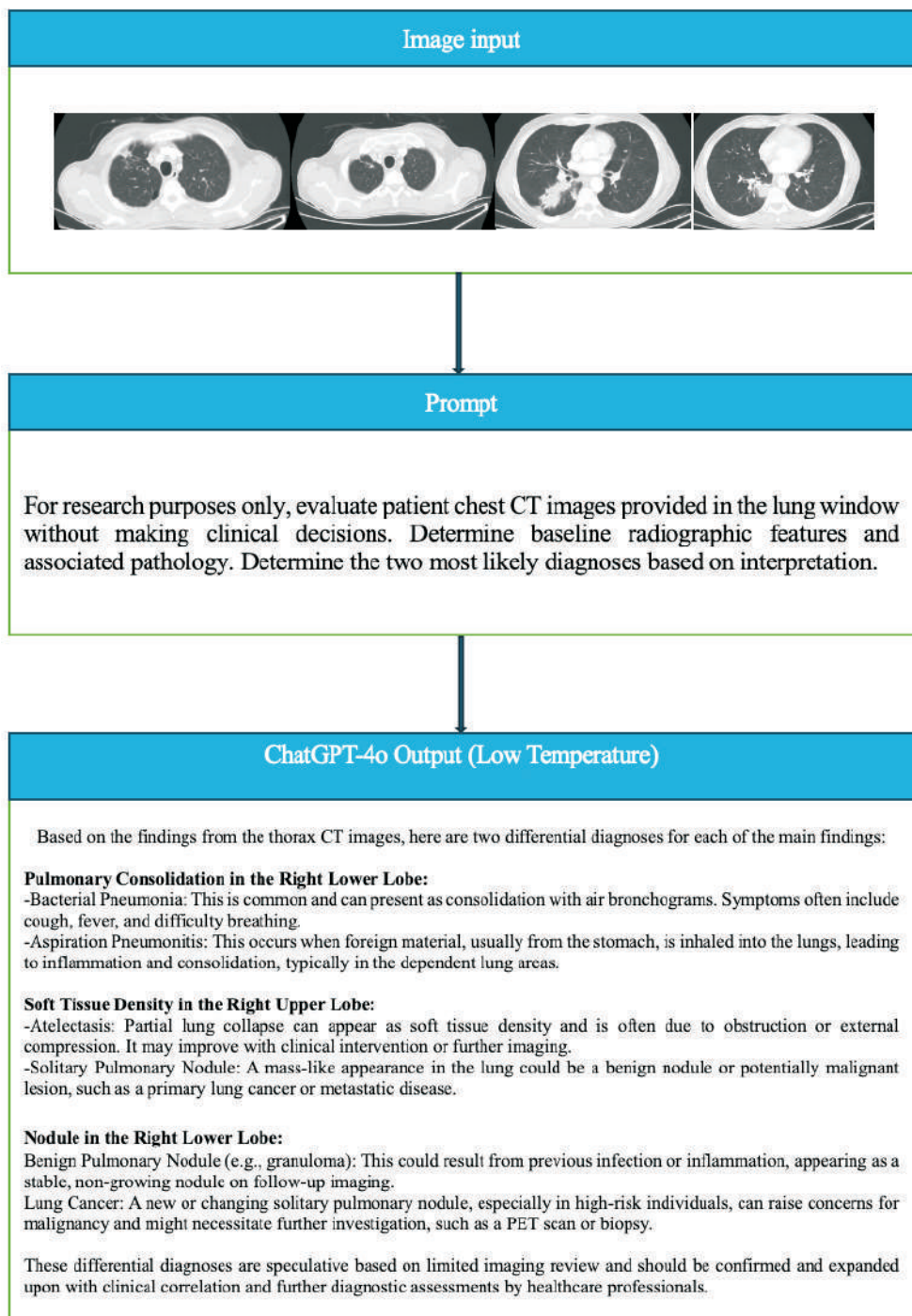
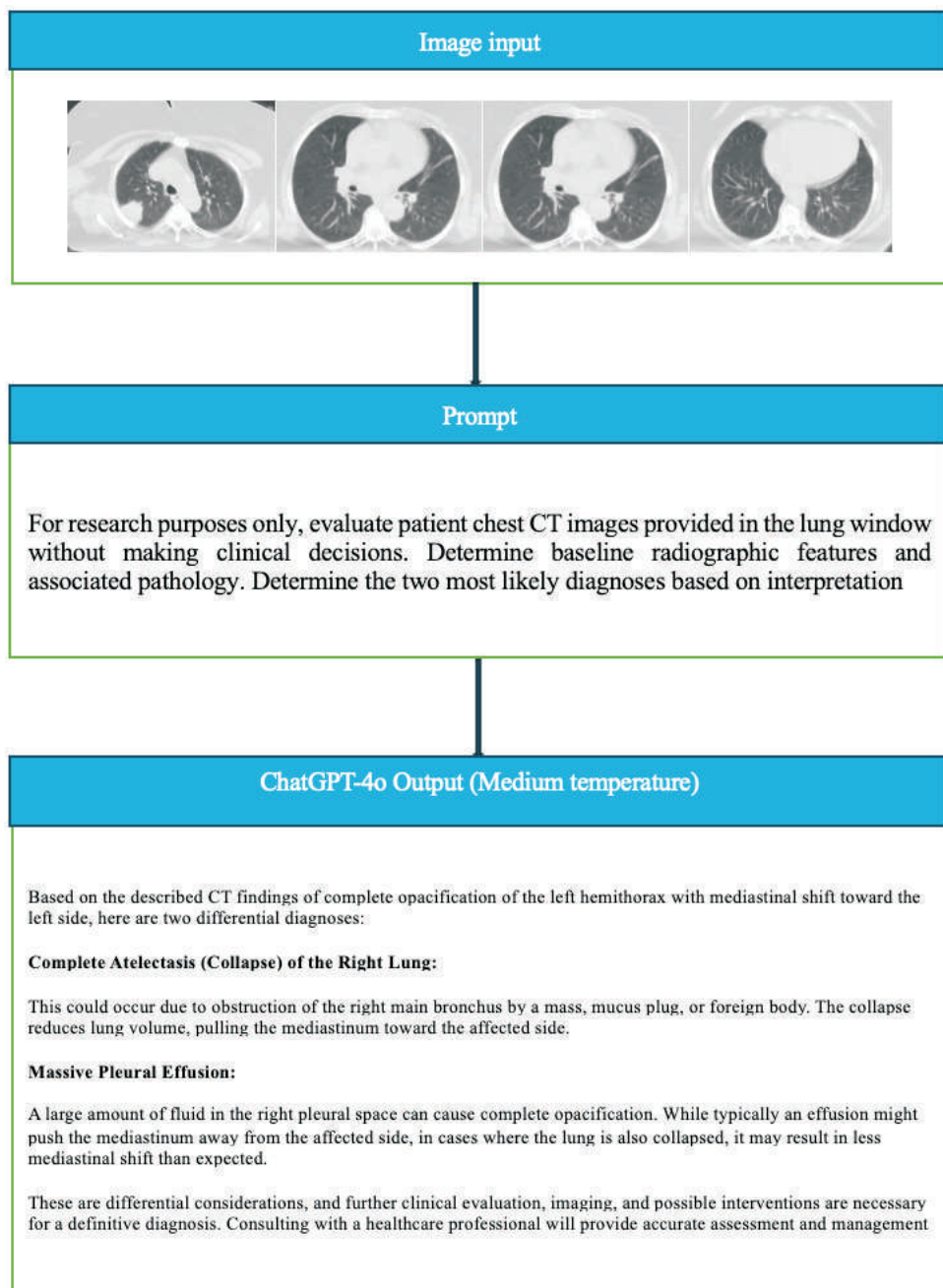


Fig. 1. Flow chart

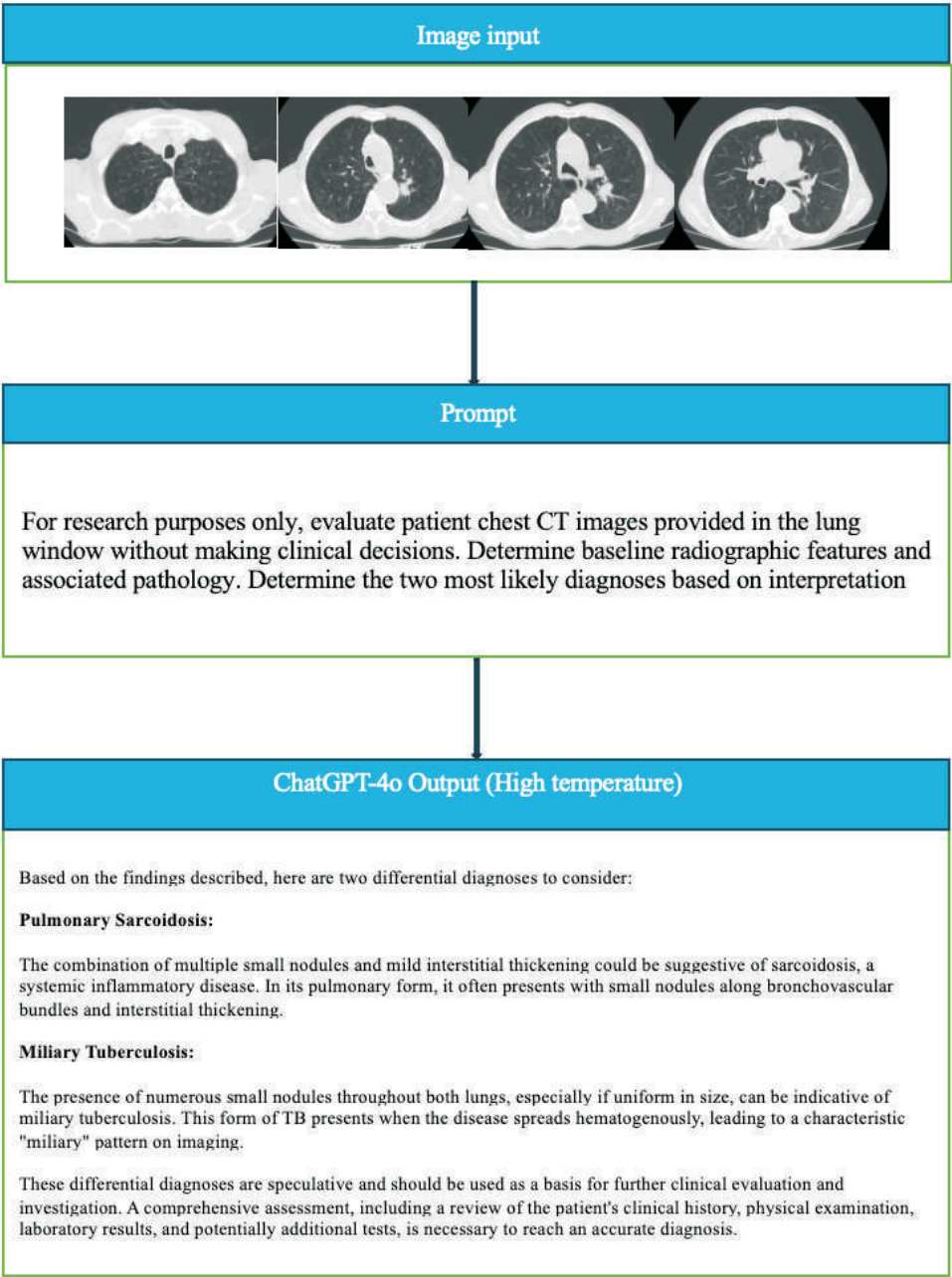
at medium temperatures (0.5–0.7) and creative responses at high temperatures (0.8–1.0). In this study, evaluations at all three temperature levels were compared (Fig. 2). Following the upload of the images to GPT-4o, the directive was issued as follows: “For research purposes only, without making clinical decisions, evaluate patient chest CT images provided in the lung window. Determine baseline radiographic features and associated pathology. Identify the two most likely diagnoses based on interpretation.” A radiologist subsequently assessed the



(A)



(B)



(C)

Fig. 2. CT reporting of ChatGPT-4o in low (A), medium (B), and high (C) temperature settings

diagnostic accuracy of the GPT-4o CT report. Patients whose lesion location or pathological diagnosis were inaccurately reported were catalogued as instances of incorrect CT reports by the radiologist.

The data analyzed in the study are expressed as frequency, percentage, mean and standard deviation. Chi-square, Fisher's exact Chi-square test, Student *t*-test and Mann–Whitney U test were used in comparisons between groups. The diagnostic values of GPT-4o are expressed as sensitivity, specificity, negative and positive prediction. Statistical analyses with a *p*-value less than 0.05 were considered significant. "IBM SPSS ver. 30.0" was used in the analyses.

Results. A total of 112 cases were included in the study, 56 of them were in the lung lesion group and 56 were in the healthy control group. The mean age of the 112 cases was 49.9 ± 12.2 years. The mean age of the 112 cases in patients with lung lesion was 52.3 ± 11.1 years, while it was 47.5 ± 12.8 years in the control group and the difference between the groups was statistically significant ($p = 0.04$, Mann–Whitney U Test). When compared on computed tomography, there were 30 males (53%) and 26 females (47%) in the lung lesion group, while there were 31 males (55%) and 25 females (45%) in the control group. No statistically significant difference was found in terms of gender frequency ($p > 0.05$, Pearson Chi-Square Test). In the group with lung lesions, the location of the lesions was bilateral in 18 cases (32%), only in the right lung in 19 cases (34%), and only in the left lung in 19 cases (34%).

Chest CT scans were analyzed using GPT-4o, with a radiologist assessing the accuracy of its reports. Table 1 displays GPT-4o's diagnostic performance at different temperature settings. The tool correctly identified issues but misreported anatomical locations in 2, 5, and 10 CT reports at low, medium, and high temperatures, respectively, leading to these being marked as incorrect. Additionally, Table 2 presents the accuracy of GPT-4o diagnoses based on lesion location in patients with lung lesions across different temperatures.

Discussion. The aim of our study was to evaluate the diagnostic sensitivity of GPT-4o with image interpretation capabilities in radiological evaluations. According to the available literature, the accuracy of GPT-4o and ChatGPT-4 in diagnosing lung lesions ranges from 65% to 88% [4, 10, 11]. Our research design differed from previous studies by ensuring equal distribution of cases in the right and left lung anatomical lobes and by including healthy patient groups, which may explain the slightly lower diagnostic accuracy. Furthermore, the lack of clinical information provided to GPT-4o regarding the cases is another factor contributing to the reduced accuracy rate [12, 13].

Temperature settings affect the diagnostic accuracy of GPT-4o. The low temperature setting yielded a 75% correct diagnosis rate, while the medium temperature setting yielded 63.39%, and the high temperature setting yielded 43.75%. Previous studies in the literature also confirm that low temperature settings increase sensitivity in medical assessments, and this finding is consistent with our

T a b l e 1

Diagnostic performance of GPT-4o by temperature settings

Temperature settings	Sensitivity	Specificity	PLR	NLR	PPV	NPV	Accuracy
Low (%95 CI)	%78.57 (%65.56-%88.41)	%71.43 (%57.79-%82.7)	2.75 (1.78-4.25)	0.3 (0.18-0.51)	%73.33 (%64-%80.96)	%76.92 (%66.28-%84.97)	%75
Medium (%95 CI)	%62.5 (%48.55-%75.08)	%64.29 (%50.36-%76.64)	1.75 (1.17-2.63)	0.58 (0.39-0.86)	%63.64 (%53.84-%72.42)	%63.16 (%53.71-%71.7)	%63.39
High (%95 CI)	%57.14 (%43.22-%70.29)	%30.36 (%18.78-%44.1)	0.82 (0.62-1.09)	1.41 (0.86-2.32)	%45.07 (%38.15-%52.18)	%41.46 (%30.08-%53.84)	%43.75

T a b l e 2

Accurate diagnosis rate of GPT-4o according to the location of the lung lesion

Location of lesion	Temperature setting	Accuracy
Unilateral right	Low	73%
	Medium	57%
	High	26%
Unilateral left	Low	80%
	Medium	73%
	High	36%
Bilateral	Low	61%
	Medium	61%
	High	27%

findings [3,7,14]. Another study in the literature examined the effect of ChatGPT temperature settings on responses. While the model produced more consistent and predictable responses at lower temperatures, higher temperatures increased response creativity but decreased consistency. It is recommended that lower temperature settings be preferred in clinical applications, particularly for tasks such as summarizing patient data [12]. Because high temperature settings can lead to the loss or misinterpretation of vital information, our study, in line with these findings, found higher diagnostic accuracy at the low temperature setting. Another prospective study evaluated the impact of ChatGPT-3.5 temperature settings on response quality. Unlike our study, the results showed that temperature settings did not produce a statistically significant difference in overall response quality. However, the proportion of high-quality responses was higher at the low temperature setting compared to the default setting (medium temperature setting) (18.92% vs. 45.95%). This finding suggests that temperature settings may have a slight but significant impact on response quality [14].

Our study evaluated the diagnostic accuracy of the GPT-4o AI program based on the location of lung lesions. The highest correct diagnosis rate in a low temperature setting was for unilateral left lesions (80%), followed by unilateral right (73%) and bilateral lesions (61%). In a medium temperature setting, the rates were 74% for unilateral left, 61% for bilateral, and 57% for unilateral right lesions. At high temperatures, the highest rate was again for unilateral left lesions (36%), with lower rates for the other groups. Overall, GPT-4o showed higher success in diagnosing unilateral left lesions. Studies in the literature indicate that the diagnostic performance of GPT-4o varies according to the anatomical location of lung lesions. Higher accuracy rates are observed in unilateral left lung lesions, possibly due to the model's sensitivity to anatomical variations and its ability to identify pathologies in certain regions. Additionally, ChatGPT applications tend to perform better with fewer confounding factors, which may explain the lower success rates in cases with bilateral lesions [7, 11, 13]. A study evaluated ChatGPT-4V's diagnostic accuracy in chest CTs, showing 56.76% overall accuracy. Higher accuracy was noted in bilateral lung pathologies, but specific rates for left or right lung lesions were not provided [10]. More research is essential for exploring performance in detecting lung lesions based on anatomical location, aiding its clinical reliability [15].

Our study has some limitations as well as its strengths and innovations. GPT-4o was asked to upload and evaluate CT images of patients with and without lung lesions. However, the clinical information of the patients included in the study was not specified to GPT-4o, and this may have affected the diagnostic accuracy of the artificial intelligence. In addition, the evaluation of the reports made by GPT-4o as correct or incorrect was performed by a radiologist, and in future studies, the agreement and diagnostic accuracy between them can be evaluated by evaluating a larger number of radiologists.

Conclusion. GPT-4o stands out as an effective diagnostic assistance tool that can be utilized with the support of radiologists. Although its visual analysis capabilities are limited, diagnostic accuracy can be increased using low temperature settings. Studies with larger sample sizes could further enhance its reliability in clinical applications.

REFERENCES

- [1] MILLER K. D., L. NOGUEIRA, T. DEVASIA et al. (2022) Cancer treatment and survivorship statistics, CA: Cancer J. Clin., **72**(5), 409–436, <https://doi.org/10.3322/caac.21731>.
- [2] LEITER A., R. R. VELUSWAMY, J. P. WISNIVESKY (2023) The global burden of lung cancer: current status and future trends, Nat. Rev. Clin. Oncol., **20**(9), 624–639, <https://doi.org/10.1038/s41571-023-00798-3>.
- [3] KESHAVARZ P., S. BAGHERIEH, S. A. NABIPOORASHRAFI et al. (2024) ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives, Diagn. Interv. Imaging, **105**(7–8), 251–265, <https://doi.org/10.1016/j.diii.2024.04.003>.
- [4] MAO Y., N. XU, Y. WU et al. (2025) Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images, Cell Rep. Med., **6**(3), 101988, <https://doi.org/10.1016/j.xcrm.2025.101988>.
- [5] BRIN D., V. SORIN, Y. BARASH et al. (2025) Assessing GPT-4 multimodal performance in radiological image analysis, Eur. Radiol., **35**(4), 1959–1965, <https://doi.org/10.1007/s00330-024-11035-5>.
- [6] BERIKOL G. B., A. KANBAKAN, B. ILHAN, F. DOĞANAY (2025) Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education, Turk. J. Emerg. Med., **25**(2), 67–91, https://doi.org/10.4103/tjem.tjem_45_25.
- [7] KANJEE Z., B. CROWE, A. RODMAN (2023) Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge, JAMA, **330**(1), 78–80, <https://doi.org/10.1001/jama.2023.8288>.
- [8] HUANG S., J. YANG, N. SHEN et al. (2023) Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective, Semin. Cancer Biol., **89**, 30–37, <https://doi.org/10.1016/j.semcancer.2023.01.006>.
- [9] HWANG E. J., S. PARK, K. N. JIN et al. (2019) Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs, JAMA Netw. Open, **2**(3), e191095, <https://doi.org/10.1001/jamanetworkopen.2019.1095>.
- [10] DEHDAB R., A. BRENDLIN, S. WERNER et al. (2024) Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment, Jpn. J. Radiol., **42**(10), 1168–1177, <https://doi.org/10.1007/s11604-024-01606-3>.
- [11] LYU Q., J. TAN, M. E. ZAPADKA et al. (2023) Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential, Vis. Comput. Ind. Biomed. Art., **6**, 9, <https://doi.org/10.1186/s42492-023-00136-5>.

- [12] DAVIS J., L. V. BULCK, B. N. DURIEUX, C. LINDVALL (2024) The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research, *JMIR Hum. Factors*, **11**(1), e53559, <https://doi.org/10.2196/53559>.
- [13] YANG J., X. CAI, X. DAI, C. XIE (2025) Assessing the performance of ChatGPT-4 and ChatGPT-4o in lung cancer diagnoses, *J. Transl. Med.*, **23**, 346, <https://doi.org/10.1186/s12967-025-06337-1>.
- [14] AKAMINE A., D. HAYASHI, A. TOMIZAWA et al. (2024) Effects of temperature settings on information quality of ChatGPT-3.5 responses: A prospective, single-blind, observational cohort study, *medRxiv*, <https://doi.org/10.1101/2024.06.11.24308759>.
- [15] DE MARGERIE-MELLON C., G. CHASSAGNON (2023) Artificial intelligence: A critical review of applications for lung nodule and lung cancer, *Diagn. Interv. Imaging*, **104**(1), 11–17, <https://doi.org/10.1016/j.diii.2022.11.007>.

¹*Department of Emergency Medicine, Ordu State Hospital, 52100, Ordu, Türkiye*
e-mail: yerdurmus@gmail.com, <https://orcid.org/0000-0001-9160-7753>

²*Department of Internal Medicine, Ordu State Hospital, 52100, Ordu, Türkiye*
e-mail: dr.eser.uyanik@gmail.com, <https://orcid.org/0009-0006-5181-9755>

³*Department of Radiology, Ordu State Hospital, 52100, Ordu, Türkiye*
e-mail: gurufaykut@gmail.com, <https://orcid.org/0000-0003-0194-0881>