JOMDI

Journal of Medical and Dental Investigations

J Med Dent Invest 2025; 6: e250132
https://doi.org/10.5577/jomdi.e250132

Original Article Open Access

Evaluation of the usability of ChatGPT-4 Pro and Gemini 2.5 Pro in patient education about brain tumors

Umut Ogün Mutlucan[1], Cihan Bedel[2], Fatih Selvi[2], Ökkeş Zortuk[3], Cezmi Çağrı Türk[1]

[1] Health Science University, Antalya Training and Research Hospital, Department of Neurosurgery, Antalya, Türkiye
[2] Health Science University, Antalya Training and Research Hospital, Department of Emergency Medicine, Antalya, Türkiye
[3] Hatay Research and Training Hospital, Department of Emergency Medicine, Hatay, Türkiye

Received: 31 October 2025
Accepted: 14 December 2025
Published: 25 December 2025

Correspondence:

Dr. Umut Ogün MUTLUCAN
Health Science University Antalya Training and Research Hospital, Department of Neurosurgery, Antalya, Türkiye.
E-mail: umutlucan@gmail.com

How to cite this article:

Mutlucan UO, Bedel C, Selvi F, Zortuk Ö, Türk CÇ. Evaluation of the usability of ChatGPT-4 and Gemini 2.5 Pro in patient education about brain tumors. J Med Dent Invest 2025;6:e250132. https://doi.org/10.5577/jomdi.e250132

Abstract

Aim: The aim of this study was to assess the reliability of ChatGPT-4 Pro and Gemini 2.5 Pro chatbots through a systematic evaluation of the responses provided by neurosurgical specialists to patients' inquiries regarding brain tumors.

Methods: The present study was conducted using artificial intelligence programs, and there is no relationship between the authors and the AI companies and sites associated with the study. The final tally revealed that a total of 56 frequently asked questions were identified. The present study will examine the ChatGPT-4 Pro and Gemini 2.5 Pro. The responses furnished by both artificial intelligence models were produced in Turkish and subsequently assessed by two independent neurosurgeons. The evaluation of the responses was conducted by two independent evaluators, who assigned scores without the knowledge of each other's evaluations. In the event that two neurosurgeons assigned the same score to a given response, it was accepted as final. In instances of discordance, a collaborative discussion was initiated, culminating in the determination and documentation of a consensus score.

Results: The distribution of these questions is illustrated in Table 2. The mean GPT score for anatomy questions was 4.25 ± 0.88, while the mean GEMINI score was 4.50 ± 0.53 (p = 0.282). For inquiries pertaining to general questions, the mean GPT score was 4.43 ± 0.81, while the mean GEMINI score was 4.38 ± 0.81 (p = 0.500). Inquiries pertaining to prognostication and daily living activities revealed a mean GPT score of 5.00 ± 0.00, accompanied by a mean GEMINI score of 4.57 ± 0.78 (p = 0.100). In the treatment questions, the mean GPT score was 4.63 ± 0.67, while the mean GEMINI score was 4.36 ± 0.67 (p = 0.138). Figure 1 presents a comparative analysis categorized by inquiry group. A comparison of the mean scores revealed that GPT and GEMINI exhibited similar performance, with mean scores of 4.54 and 4.44, respectively.

Conclusion: The present study demonstrates that large language model (LLM) technologies, including ChatGPT-4 Pro and Gemini 2.5 Pro, exhibit considerable promise in the provision of information and guidance to patients. Furthermore, the investigation revealed that artificial intelligence models do not demonstrate a substantial degree of superiority in the education of patients regarding brain tumors.

Keywords: Artificial intelligence, ChatGPT-4 Pro, Gemini 2.5 Pro, chatbot, brain, tumor

OPEN ACCESS

This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0) (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Artificial intelligence (AI) is a multidisciplinary field of enquiry focused on the creation of systems that simulate human intelligence. These systems can acquire knowledge, solve problems, and make decisions (1). AI is a multidisciplinary field that incorporates elements of computer science, psychology, and engineering. The development of AI systems capable of performing tasks autonomously or semi-autonomously constitutes a significant area of research. Significant advances have been made in the field, particularly in machine learning and neural networks. These advancements have enabled the application of AI in diverse areas, including healthcare, robotics, and natural language processing (2).

Chat Generative Pre-trained Transformer (ChatGPT) and Google Gemini represent two advanced AI models that have been evaluated across a range of domains, including medical and business applications. It is evident that both models exhibit distinct strengths and limitations, which, in turn, influence their effectiveness across different contexts. ChatGPT, developed by OpenAI, is characterized by its conversational fluency and adaptability. Conversely, Google Gemini demonstrates proficiency in contextual variety and data-driven decision-making (2-4).

Individuals afflicted with brain tumors frequently have a multitude of inquiries and apprehensions that traverse medical, psychological, and practical domains. These inquiries may encompass a wide spectrum, ranging from pressing medical concerns to long-term implications for quality of life. Addressing these questions effectively requires a multidisciplinary approach that integrates clinical nurse specialists, question prompt lists, and psycho-oncological support. Inquiries regarding the consequences of the illness on day-to-day activities and professional obligations are a common occurrence (5, 6).

Research has demonstrated that patients derive substantial benefit from a robust support network comprising family members, caregivers, and healthcare professionals. In addressing these concerns, it is imperative to acknowledge the uniqueness of each patient's experience. It is acknowledged that patients may assign varying degrees of importance to different aspects of their care and that their requirements may undergo change over time. It is evident that patients frequently utilize online medical information to expand their understanding of their illnesses (5-7). Given the nascent stage of research in this area, the accuracy and reliability of artificial intelligence in this regard remain to be fully elucidated. The subsequent text is designed to offer a thorough and comprehensive overview of the subject matter.

The aim of this study was to assess the reliability of ChatGPT-4 Pro and Gemini 2.5 Pro chatbots through a systematic evaluation of the responses provided by neurosurgical specialists to patients' inquiries about brain tumors.

Materials and Methods

This study used artificial intelligence programs, and the authors have no relationship with the AI companies or sites associated with the study. The authors conducted a comprehensive review of neurosurgical websites, including the National Cancer Institute (www.cancer.gov), the American Association of Neurological Surgeons, educational health platforms for patients, and public forums such as Quora. The goal was to identify frequently asked questions about brain tumors and their treatment. The authors then conducted a thorough review of the questions, excluding duplicates, irrelevant questions, vague questions, and non-subjective questions. The final tally revealed 56 frequently asked questions. These questions underwent linguistic revision and were systematically categorized into thematic groups. The following categories were considered: anatomical/structural; treatment and interventions; daily life; causes and risk factors; prognosis/recovery; diagnosis and imaging; and other/general.

Each of the finalized questions was translated into Turkish and posed individually to two artificial intelligence models: ChatGPT-4 Pro and Gemini 2.5 Pro. The responses furnished by both artificial intelligence models were produced in Turkish and subsequently assessed by two independent neurosurgeons.

Responses were scored based on the following criteria:

- **1 point**: Entirely incorrect or irrelevant
- **2 points:** A mix of correct and incorrect/misleading information
- **3 points:** Correct but incomplete
- **4 points:** Academically accurate but less accessible to the general public
- **5 points:** Accurate and clearly understandable by laypersons

The evaluation of the responses was conducted by two independent evaluators, who assigned scores without the knowledge of each other's evaluations. If two neurosurgeons assigned the same score to a given response, the score was accepted as final. In instances of discordance, a collaborative discussion was initiated, culminating in the determination and documentation of a consensus score.

Statistical analysis

The statistical analysis of the study's findings was conducted using SPSS software (Version 27.0; IBM Inc., Armonk, NY, US). GraphPad Prism version 10.2 was used to generate the graphs, and the data were subsequently analyzed. Categorical data were defined as percentages and frequencies. The numerical data were defined, and a distribution analysis was performed. Data sets that conformed to a normal distribution were defined as

mean ± standard deviation (SD). The interobserver test was used to assess interrater consistency in the study. In the analysis of the relationship between numerical variables, the paired t-test was employed for data that fit a normal distribution. In this study, data were deemed significant only if the p-value was <0.05.

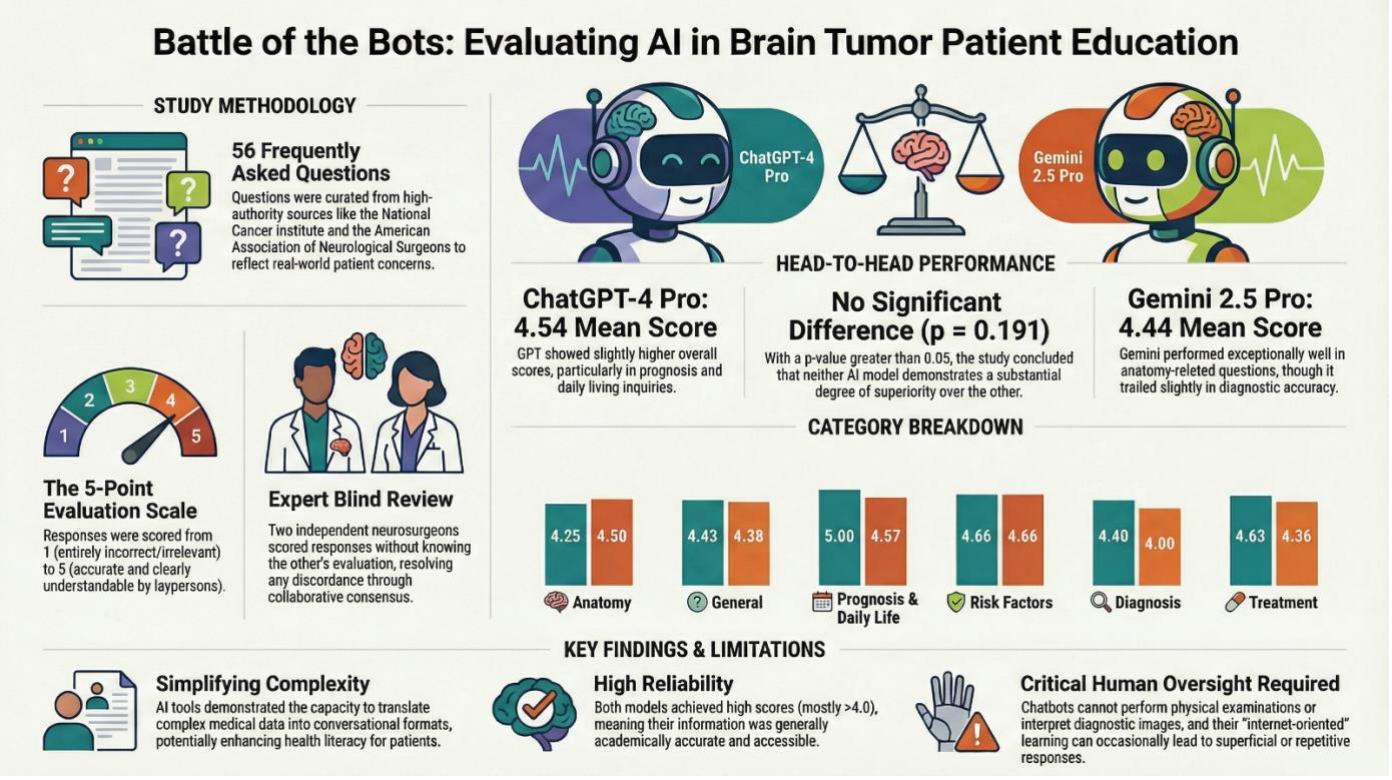

Figure 1. **Comparison by question groups.**

Results

In the study, the answers provided by LLMs were evaluated by two different experts. The evaluation analysis performed by the experts is presented in Table 1. LLMs were presented with a total of 56 questions concerning the phenomenon of brain corruption, which were grouped into six distinct categories. The distribution of these questions is illustrated in Table 2. The mean GPT score for anatomy questions was 4.25 ± 0.88, while the mean GEMINI score was 4.50 ± 0.53 (p = 0.282). For general inquiries, the mean GPT score was 4.43 ± 0.81, and the mean GEMINI score was 4.38 ± 0.81 (p = 0.500). Inquiries regarding prognostication and daily living activities revealed a mean GPT score of 5.00 ± 0.00, accompanied by a mean GEMINI score of 4.57 ± 0.78 (p = 0.100). In the context of risk assessment, the mean GPT score was determined to be 4.66±0.50, while the mean GEMINI score exhibited a similar value of 4.66±0.50 (p = 0.50). Regarding diagnostic questions, the mean score for GPT was 4.40 ± 0.89, whereas the mean score for GEMINI was 4.00 ± 1.00 (p = 0.324). In the treatment questions, the mean GPT score was 4.63 ± 0.67, while the mean GEMINI score was 4.36 ± 0.67 (p = 0.138). Figure 1 presents a comparative analysis categorized by inquiry group. A comparison of mean scores revealed that GPT and GEMINI performed similarly, with mean scores of 4.54 and 4.44, respectively. This observation did not demonstrate a statistically significant difference between the two groups (p = 0.191). Figure 2 presents a comparison of the responses to all questions.

Table 1. Interobserver correlation.

	Mean±SD	Intraclass Correlation	p-value
GPT Observer 1	4.55±0.71	0.046	0.368
GPT Observer 2	4.21±0.75		
GEMINI Observer 1	4.44±0.71	0.530	0.003
GEMINI Observer 2	4.69±0.60		

Table 2. Distribution of questions.

	N	%
Anatomy	8	14.3%
General	16	28.6%
Prognosis, Daily life	7	12.5%
Risk	9	16.1%
Diagnosis	5	8.9%
Treatment	11	19.6%

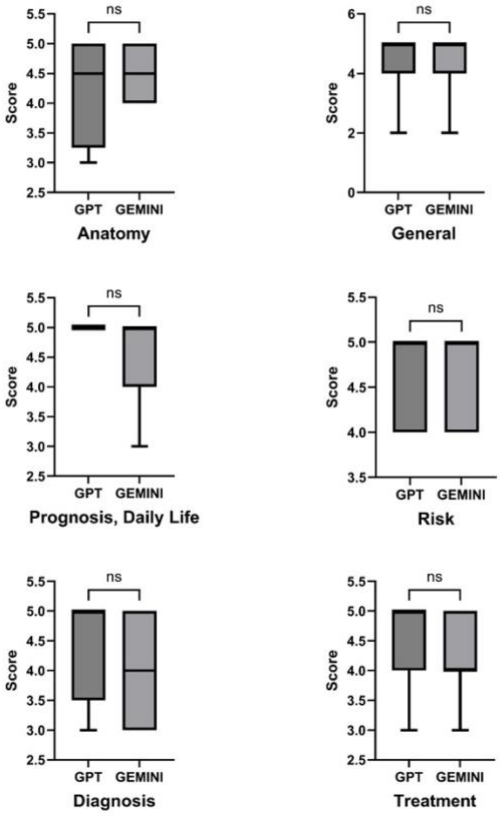

Figure 1. Comparison by question groups.

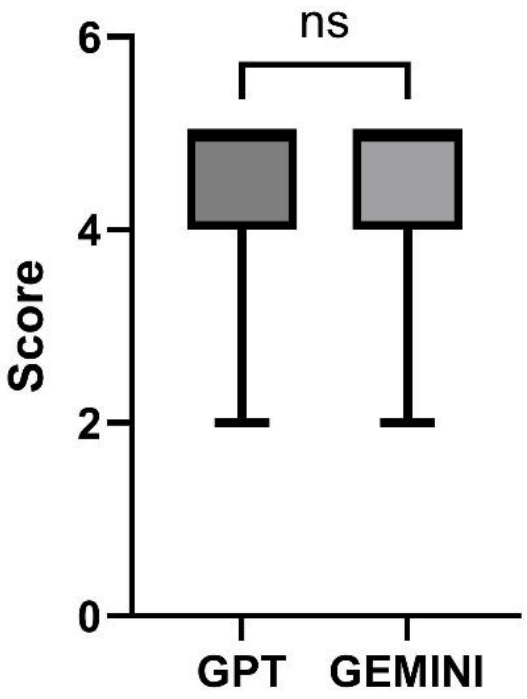

Figure 2. Comparison by total score.

Discussion

Recent studies have begun to examine the potential of AI for usability, particularly in the follow-up and treatment of various internal and surgical diseases. In the present study, the mean GPT score was 4.54 ± 0.71, whereas the mean GEMINI score was 4.44 ± 0.71. A paired-samples t-test was performed, and the result indicated that there was no significant difference between the two groups (p = 0.191).

The integration of AI into patient education is a growing trend, offering innovative solutions to enhance understanding and engagement across various medical fields. The potential of AI to personalize educational content, enhance accessibility, and facilitate interactive learning experiences is profoundly impacting the manner in which patients receive and process health-related information. Nevertheless, challenges such as accuracy, readability, and ethical considerations remain critical in the deployment of AI in patient education (8,9). The capacity of artificial intelligence to customize educational content according to the specific requirements of individual patients is a notable development. The integration of personalized recommendations and interactive elements, such as chatbots and virtual assistants, has been demonstrated to enhance engagement and motivation. Within the domain of dentistry, the integration of AI has the potential to enhance patient education through personalized materials and virtual consultations, thereby optimizing patient comprehension and experience (10,

11). In our study, the mean GPT score for questions concerning diagnosis was 4.40 ± 0.89, whereas the mean GEMINI score was 4.00 ± 1.00 (p = 0.324). In the treatment questions, the mean GPT score was 4.63 ± 0.67, while the mean GEMINI score was 4.36 ± 0.67 (p = 0.138).

While AI-generated materials have exhibited potential in enhancing readability, traditional patient information leaflets frequently demonstrate superiority in readability metrics. AI tools such as ChatGPT have the capacity to simplify medical information by rendering it in more conversational formats, with the potential to enhance health literacy. The accuracy of AI-generated educational materials is comparable to that of materials created by healthcare professionals, though occasional inaccuracies, particularly in complex medical topics, have been noted (11-13). Pharmacists have expressed a reasonable degree of confidence in the clinical accuracy of AI-generated materials, emphasizing the necessity for additional validation. Concerns regarding the overreliance on AI, the financial implications of implementation, patient acceptance, privacy, and regulatory compliance persist. Ensuring the ethical use and equitable access to AI-driven educational tools is imperative, particularly in the context of diverse patient populations (14). The capacity of artificial intelligence to monitor patient progress, analyze feedback, and optimize administrative processes has the potential to enhance patient education and support. It is recommended that future research endeavors concentrate on the development of standardized approaches for the evaluation of AI's impact on patient education and the exploration of its integration into clinical practice. While there is considerable potential for artificial intelligence (AI) to transform patient education, it is vital to strike a balance between innovation and evidence-based healthcare practices (13-15). In our study on the treatment questions, the mean GPT score was 4.63 ± 0.67, while the mean GEMINI score was 4.36 ± 0.67 (p = 0.138).

The integration of ChatGPT and its successor, GPT-4, into the domain of neurosurgery represents a development that carries with it both opportunities and challenges. These AI models have demonstrated considerable potential in a variety of applications, including the creation of discharge summaries, operative reports, and educational materials. Furthermore, they have demonstrated a high level of proficiency in neurosurgical board examinations. Nevertheless, limitations such as a paucity of depth in responses, an incapacity to perform physical examinations, and ethical concerns persist as significant hurdles. Subsequent sections delve into the specific roles and limitations of ChatGPT within the domain of neurosurgery (15, 16). Recent studies have demonstrated that ChatGPT can significantly reduce the time required to prepare neurosurgical discharge summaries and operative reports when compared to traditional methods. Moreover, it has been demonstrated to exhibit a high degree of factual accuracy. Educational Tool: GPT-4 has exhibited superior performance in board-style questions, surpassing the performance of medical students and neurosurgical residents, suggesting its potential as a valuable educational resource for neurosurgical training (17).

GPT-4 has exhibited a high degree of accuracy in the diagnosis and triaging of neurosurgical scenarios, comparable to that of senior neurosurgical residents. A recent analysis of ChatGPT's responses in the domain of neurosurgery reveals a tendency towards superficiality and repetition. This finding indicates a deficiency in the depth and personalization that are crucial for effective patient care. ChatGPT exhibits deficiencies in tasks requiring physical examination or interpretation of diagnostic images, which are of paramount importance in the domain of neurosurgical practice (18). While ChatGPT and GPT-4 have demonstrated potential in enhancing neurosurgical practice and education, their current limitations necessitate a cautious integration into clinical workflows. The potential for artificial intelligence (AI) to augment human intelligence in the domain of neurosurgery is significant. However, this potential is contingent upon the careful consideration of the ethical implications of such practices, as well as the continuous improvement of AI models to address existing shortcomings (19-21). We found that A comparison of the mean scores revealed that GPT and GEMINI exhibited similar performance, with mean scores of 4.54 and 4.44, respectively. This observation did not demonstrate a statistically significant difference between the two groups (p = 0.191).

The study's limitations can be attributed to the internet-oriented nature of the responses provided by the chatbots. Recent findings have revealed that these responses can result in confusion due to the implementation of erroneous learning algorithms. Despite the efforts of artificial intelligence companies to mitigate this issue, it is important to acknowledge the possibility of encountering this problem in such chatbots. In the present study, as outlined in the appendix, patient questions were predominantly utilized and evaluated. It is an intriguing prospect to contemplate the potential applications of artificial intelligence programs in addressing patient queries.

Conclusion

The findings of the present study demonstrate that large language models (LLMs), including ChatGPT and Google Gemini, show considerable promise in providing information and guidance to patients. Furthermore, the investigation revealed that artificial intelligence models do not demonstrate a substantial degree of superiority in the education of patients regarding brain tumors.

Disclosures

Peer-review: Externally peer-reviewed.

Author Contributions: Concept — U.O.M.; Design — U.O.M.; Supervision — C.B., F.S.; Materials — U.O.M.; Data collection &/or processing — Ö.Z., C.Ç.T.; Analysis and/or interpretation — Ö.Z., C.Ç.T.; Literature search — F.S., Ö.Z., C.Ç.T.; Writing — U.O.M.; Critical review — Ö.Z., C.Ç.T.

Conflict of Interest: There is no any conflict of interest.

Funding: There is no any funding.

References

1. Cè M, Irmici G, Foschini C, Danesini GM, Falsitta LV, Serio ML, et al. Artificial intelligence in brain tumor imaging: a step toward personalized medicine. Curr Oncol. 2023;30(3):2673-2701. https://doi.org/10.3390/curroncol30030203
2. Huang J, Shlobin NA, Lam SK, DeCuypere M. Artificial intelligence applications in pediatric brain tumor imaging: a systematic review. World Neurosurg. 2022;157:99-105. https://doi.org/10.1016/j.wneu.2021.10.068
3. Taşyürek M, Adıgüzel Ö, Gündoğar M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. Int Dent Res. 2025;15(3):123-135. https://doi.org/10.5577/intdentres.662
4. Afridi M, Jain A, Aboian M, Payabvash S. Brain tumor imaging: applications of artificial intelligence. Semin Ultrasound CT MR. 2022;43(2):153-169. https://doi.org/10.1053/j.sult.2022.02.005
5. Luo J, Pan M, Mo K, Mao Y, Zou D. Emerging role of artificial intelligence in diagnosis, classification and clinical management of glioma. Semin Cancer Biol. 2023;91:110-123. https://doi.org/10.1016/j.semcancer.2023.03.006
6. Raghavendra U, Gudigar A, Paul A, Goutham TS, Inamdar MA, Hegde A, et al. Brain tumor detection and screening using artificial intelligence techniques: current trends and future perspectives. Comput Biol Med. 2023;163:107063. https://doi.org/10.1016/j.compbiomed.2023.107063
7. Hsieh KL, Chen CY, Lo CM. Quantitative glioma grading using transformed gray-scale invariant textures of MRI. Comput Biol Med. 2017;83:102-108. https://doi.org/10.1016/j.compbiomed.2017.02.012
8. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045-1057. https://doi.org/10.1007/s10278-013-9622-7
9. Gutta S, Acharya J, Shiroishi MS, Hwang D, Nayak KS. Improved glioma grading using deep convolutional neural networks. AJNR Am J Neuroradiol. 2021;42(2):233-239. https://doi.org/10.3174/ajnr.A6882
10. Wang G, Li W, Ourselin S, Vercauteren T. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. Front Comput Neurosci. 2019;13:56. https://doi.org/10.3389/fncom.2019.00056
11. Kabir Anaraki A, Ayati M, Kazemi F. Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. Biocybern Biomed Eng. 2019;39(1):63-74. https://doi.org/10.1016/j.bbe.2018.10.004
12. Yang Y, Yan LF, Zhang X, Han Y, Nan HY, Hu YC, et al. Glioma grading on conventional MR images: a deep learning study with transfer learning. Front Neurosci. 2018;12:804. https://doi.org/10.3389/fnins.2018.00804
13. Ertosun MG, Rubin DL. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. AMIA Annu Symp Proc. 2015;2015:1899-1908.
14. Zhuge Y, Ning H, Mathen P, Cheng JY, Krauze AV, Camphausen K, et al. Automated glioma grading on conventional MRI images using deep convolutional neural networks. Med Phys. 2020;47(7):3044-3053. https://doi.org/10.1002/mp.14168
15. Eichinger P, Alberts E, Delbridge C, Trebeschi S, Valentinitsch A, Bette S, et al. Diffusion tensor image features predict IDH genotype in newly diagnosed WHO grade II/III gliomas. Sci Rep. 2017;7(1):13396. https://doi.org/10.1038/s41598-017-13679-4
16. Lohmann P, Lerche C, Bauer EK, Steger J, Stoffels G, Filss CP, et al. Predicting IDH genotype in gliomas using FET PET radiomics. Sci Rep. 2018;8(1):13328. https://doi.org/10.1038/s41598-018-31806-7
17. Zaragori T, Oster J, Roch V, Hossu G, Chary-Valckenaere I, Blum A, et al. 18F-FDOPA PET for the non-invasive prediction of glioma molecular parameters: a radiomics study. J Nucl Med. 2022;63(1):147-153. https://doi.org/10.2967/jnumed.120.261545
18. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005;352(10):997-1003. https://doi.org/10.1056/NEJMoa043331
19. Mutlucan UO, Zortuk O, Bedel C, Selvi F, Türk CC. Comparison of ChatGPT and Gemini in neurosurgical evaluation questions. Sarcouncil J Med Surg. 2024;3(12):10-15.
20. Hegi ME, Liu L, Herman JG, Stupp R, Wick W, Weller M, et al. Correlation of O6-methylguanine methyltransferase (MGMT) promoter methylation with clinical outcomes in glioblastoma and clinical strategies to modulate MGMT activity. J Clin Oncol. 2008;26(25):4189-4199. https://doi.org/10.1200/JCO.2007.11.5964
21. Bedel HA, Bedel C, Selvi F, Zortuk Ö, Karancı Y. Emergency medicine assistants in the field of toxicology, comparison of ChatGPT-3.5 and GEMINI artificial intelligence systems. Acta Med Litu. 2024;31(2):294-301. https://doi.org/10.15388/Amed.2024.31.2.18