

DIAGNOSTIC ACCURACY OF LARGE LANGUAGE
MODELS FOR FICTIONAL PEDIATRIC
NEURORADIOLOGY CASES: A COMPARATIVE STUDY
WITH RADIOLOGISTS

Turay Cesur¹✉, Yasin Celal Gunes², Eren Camur³,
Gulay Cesur Cinar⁴, Goksel Tuzcu⁵, Avni Merter Keceli⁵

Received on March 2, 2026

Presented by D. Damianov, Member of BAS, on April 28, 2026

Abstract

The aim of this study was to evaluate diagnostic accuracy and differential-diagnosis quality of large language models (LLMs) for clinically realistic, text-only pediatric neuroradiology cases. This cross-sectional diagnostic accuracy study included 100 fictional pediatric neuroradiology cases composed of a brief clinical presentation and a structured text-only CT/MRI report curated by an expert pediatric neuroradiologist and a pediatricist. For each case, a reference primary diagnosis and two acceptable alternatives were pre-specified. Seven LLM variants (ChatGPT-5.2 Instant/Auto/Thinking; Gemini 3 Pro/Thinking; Claude 4.5 Opus/Opus Thinking) and three radiologists (two general radiologists; one pediatric radiologist) each provided one primary diagnosis and two differentials. Primary outcome was top-1 accuracy (exact match or acceptable alternative). Secondary outcomes were Differential Diagnosis Score (DDxScore, 1–5) and response time. Paired accuracy differences were assessed with Cochran's Q and post-hoc McNemar tests; DDxScore and response time were compared using Friedman tests with post-hoc Wilcoxon signed-rank tests and multiplicity correction. Top-1 accuracy ranged from 44–54% among radiologists (pediatric radiologist 54%) and 48–80% among LLMs (ChatGPT-5.2 Thinking/Auto and Gemini 3 Thinking 80%; Claude 4.5 Opus Thinking 76%). Overall accuracy differed across raters (Cochran's Q = 107.86, df = 9, $p < 0.001$).

<https://doi.org/10.7546/CRABS.2026.05.12>

Median DDxScores were 4.0 (IQR 2.0–4.0) for the pediatric radiologist, 3.0 (2.0–4.0) for general radiologists, and up to 5.0 (4.0–5.0) for leading thinking-mode LLMs ($p < 0.001$). All LLMs were faster than radiologists ($p < 0.001$). In text-only pediatric neuroradiology cases, several contemporary LLMs matched or exceeded radiologists in top-1 accuracy and produced high-quality differentials with substantially shorter response times. These findings support further evaluation for education and audited decision-support workflows.

Key words: large language models, artificial intelligence, pediatric neuroradiology, diagnostic accuracy, radiology education

Introduction. Pediatric neuroradiology encompasses imaging of the central and peripheral nervous systems and head/neck/spine across developmental stages, using modalities including US, CT, MRI, radiography, and angiography [1]. Despite technical advances, access to pediatric neuroimaging expertise remains uneven, underscoring the need for scalable approaches that support equitable care [2].

Large language models (LLMs) have rapidly expanded across medical applications involving complex text synthesis and explanation [3]. In radiology, LLMs have been investigated for impression generation, exam-style question answering, text-based case diagnosis, and workflow-adjacent decision support [4–8]. Broader radiology literature suggests potentially useful performance across subspecialties, but emphasizes the need for careful validation, safety assessment, and appropriate governance prior to clinical deployment [9]. Text-based neuroradiology evaluations have also reported promising results when models are provided structured clinical histories and imaging findings [4, 5, 7, 8, 10–13].

In pediatric radiology, prior work has examined LLM performance using text questions, case vignettes, bone age assessment, and visual diagnostic tasks [14–16]. However, to our knowledge, a focused, comparative evaluation of LLMs on text-only pediatric neuroradiology cases with human radiologist comparators has not been systematically reported.

Therefore, we evaluated diagnostic accuracy and differential-diagnosis quality of multiple contemporary LLMs on expert-authored fictional but clinically realistic pediatric neuroradiology cases (text-only CT/MRI reports plus brief clinical presentation). Secondary aims were to assess differential utility using an ordinal DDxScore, explore performance variation across case categories, and compare LLMs with radiologists of differing experience interpreting identical case information. We hypothesized that contemporary LLMs would demonstrate moderate-to-high top-1 accuracy and clinically useful differentials in this constrained, text-only setting.

Materials and methods. Study design and reporting. This was a case-based diagnostic accuracy study, reported with STARD 2015 principles adapted to a fictional-case design [17]. Ethics approval was not required because no patient data were used.

Case creation and reference standard. A board-certified pediatric neuro-radiologist with 20 years’ experience (A.M.K.) created 100 fictional pediatric neuroradiology cases reflecting routine reporting style, each consisting of a brief clinical vignette and a structured text-only CT/MRI report. A pediatrician with 12 years’ experience (G.C.) prepared age-appropriate clinical histories and key findings. Cases covered neoplasms, infection/inflammation, vascular lesions, congenital/developmental anomalies, and metabolic/neurodegenerative disorders. For each case, the reference standard consisted of one primary diagnosis defined at clinically meaningful specificity and two acceptable alternative diagnoses (e.g., entity-level vs. closely related subtype). Representative cases are shown in Table 1; study workflow is summarized in Fig. 1.

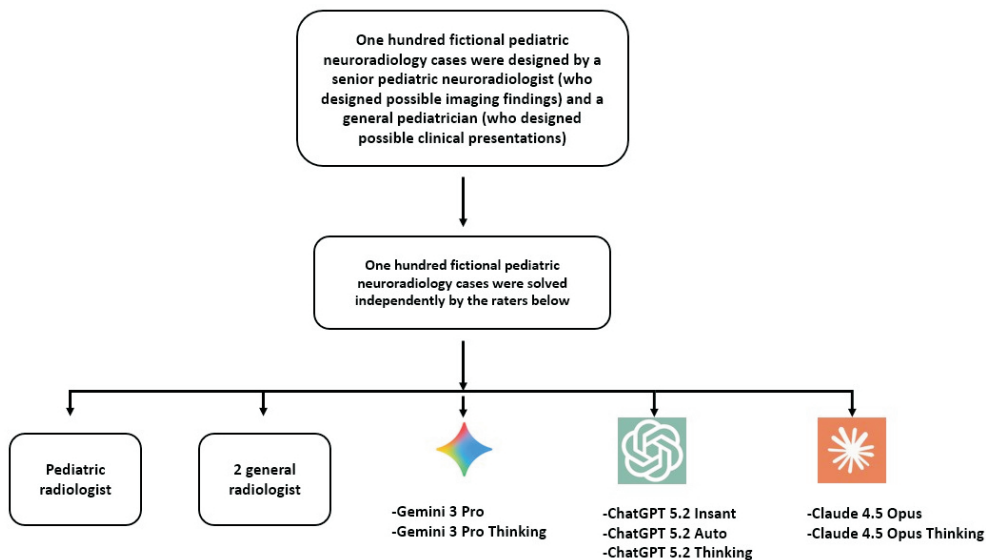


Fig. 1. Flowchart of the study

Large language models and access conditions. Seven LLMs were evaluated in December 2025: ChatGPT-5.2 (Instant/Auto/Thinking), Gemini 3 (Pro/Thinking), and Claude 4.5 Opus (standard/Thinking). All models were accessed through official web interfaces using default user-facing settings, with no external tools, retrieval, plug-ins, or parameter tuning, to reflect typical real-world use.

Prompting and data collection. Prompt structure followed prior radiology-focused LLM evaluations emphasizing explicit role framing and concise task instructions [18]. The base prompt was: “As a highly experienced professor of pediatric neuroradiology with 30 years of expertise, analyze the patient history and imaging findings and provide the most likely diagnosis and two differential diagnoses.” Each case was submitted in an independent session to minimize context carryover; no feedback or correction was provided between cases. A European

T a b l e 1

Fictional paediatric presentations, fictional radiological reports, diagnosis, and differential diagnosis sections of the cases used in our study shown with three examples

Pediatric resentation	Radiology reports	Diagnosis	Differential diagnoses
A 1-year-old boy was evaluated for delayed motor development and the absence of attempts to walk. The family also reported poor response to auditory stimuli.	Symmetric, confluent signal changes were observed in the periventricular and deep white matter, with preservation of the U-fibres. On thin-section T2-weighted sequences, punctate T2 hypointense foci were identified. The corpus callosum was of normal thickness, with no evidence of significant atrophy.	Metachromatic leukodystrophy	Pelizaeus–Merzbacher disease Krabbe disease
A 15-year-old girl presented with complaints of scalp depression, mild right-sided weakness, and recurrent headaches. The family had noticed skin retraction and a scar-like appearance for a long time.	Marked thinning of the soft tissues was observed beneath the right frontoparietal scalp. On FLAIR sequences, multiple hyperintense foci were seen in the right hemisphere involving the centrum semiovale, frontoparietal subcortical regions, and areas adjacent to the caudate nucleus, thalamus, and globus pallidus. On SWI sequences, corresponding signal dropouts consistent with calcifications were identified in the same regions. The cortical gyri appear focally thinned.	Linear scleroderma	Parry–Romberg syndrome Intracranial (neurocutaneous) infection with calcifications
A 4-month-old female infant was brought for evaluation due to early onset epileptic seizures, a whitish appearance in the eyes, and delayed head circumference growth. It was reported that she had feeding difficulties and irritability since birth.	On diffusion-weighted imaging, marked diffusion restriction was observed in the posterior limbs of the internal capsules, optic radiations, brainstem tracts, middle cerebellar peduncles, and dentate nuclei. The corpus callosum appears thin, and the dorsal brainstem is hypoplastic. On T2-weighted images, bilateral lens opacification with increased signal was noted. On MR spectroscopy, no pathological peak was identified.	ITPA-related cephalopathy	Mitochondrial disorders Krabbe disease

Board-certified general radiologist (T.C., ~ 8 years' experience) performed model submissions and recorded response times. An example input-output workflow is provided in Fig. 2.

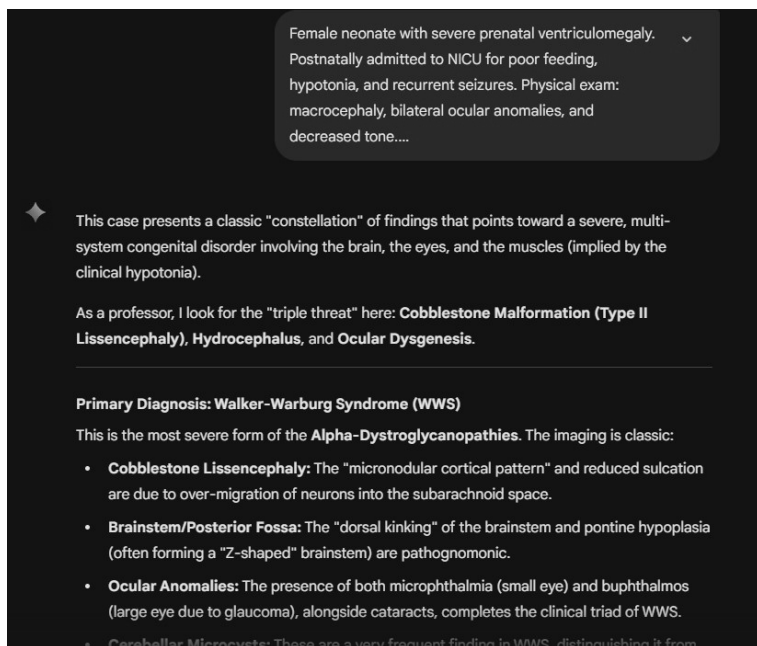


Fig. 2. An example of the response of Google Gemini 3 Pro to a fictional case involving Walker–Warburg syndrome

Human comparator group. Two European Board-certified general radiologists (8 years' experience; Y.C.G., E.Ç.) and one pediatric radiologist (15 years' experience; G.T.) interpreted the same text-only cases, blinded to reference diagnoses and LLM outputs. They provided one primary diagnosis and two differentials per case; response time was recorded. The case-creating pediatric neuroradiologist evaluated responses for accuracy and DDxScore.

Top-1 accuracy: Primary diagnosis coded as correct if it matched the reference diagnosis or one of two pre-specified acceptable alternatives; otherwise incorrect. Accuracy was the proportion correct across 100 cases.

Differential diagnosis score (DDxScore, 1–5): Adapted from prior radiology LLM evaluation frameworks that capture partial correctness and clinical usefulness [4].

- 5: primary correct; differentials complete/appropriate
- 4: primary correct; minor omissions/less relevant additions
- 3: primary incorrect; correct diagnosis present high in differential list
- 2: primary incorrect; correct diagnosis only low in differentials
- 1: correct diagnosis absent; no close alternative

Response time: Time from case submission to finalized answer (seconds).

Statistical analysis. Paired binary accuracy across ten raters was compared using Cochran's Q; significant omnibus results were followed by pairwise McNemar tests with Bonferroni-adjusted two-sided p -values ($\alpha = 0.05$). DDxScore and response times were compared using Friedman tests and post-hoc Wilcoxon signed-rank tests with Bonferroni correction. Analyses were performed using SPSS v26 (IBM).

Results. Across 100 cases, performance varied substantially (Table 2). Top-1 accuracy was 54% for the pediatric radiologist and 47%/44% for general radiologists. LLM accuracy ranged from 48% (Gemini 3 Pro) to 80% (GPT-5.2 Thinking, GPT-5.2 Auto, Gemini 3 Thinking); Claude 4.5 Opus Thinking achieved 76% and Claude 4.5 Opus 72%. Diagnostic accuracy differed significantly across raters (Cochran's $Q = 107.86$, $df = 9$, $p < 0.001$). In post-hoc comparisons, the highest-performing LLMs (80% group; Claude 4.5 Opus Thinking) outperformed each radiologist in most contrasts (many $p < 0.001$) (Table 2).

DDxScore also differed across raters (Friedman $\chi^2(9) = 192.25$, $p < 0.001$). Median (IQR) DDxScores were 4.0 (2.0–4.0) for the pediatric radiologist and 3.0 (2.0–4.0) for both general radiologists. LLM medians ranged from 3.0 (3.0–5.0) to 5.0 (4.0–5.0), with the highest median DDxScore observed for Gemini 3 Thinking and Claude 4.5 Opus Thinking (both 5.0 [4.0–5.0]). Thinking/Auto variants generally outperformed instant/pro variants (adjusted $p \leq 0.015$) (Table 2).

Response times differed significantly (Friedman $\chi^2(9) = 804.48$, $p < 0.001$). All LLMs were faster than radiologists (all adjusted $p < 0.001$). Instant-response variants were fastest overall, whereas thinking-mode variants were slower than instant variants (all adjusted $p < 0.001$), consistent with increased deliberation.

Discussion. In this text-only pediatric neuroradiology benchmark, several contemporary LLMs achieved diagnostic accuracies and differential-diagnosis quality comparable to, and in multiple comparisons exceeding, those of experienced radiologists interpreting identical structured information. To our knowledge, this represents the first focused evaluation of LLM diagnostic performance in pediatric neuroradiology using clinically realistic, expert-authored fictional reports with human comparators.

Our results extend earlier radiology literature suggesting that LLMs can perform strongly on structured text-based diagnostic tasks across subspecialties [4–9] and align with neuroradiology vignette studies showing promising diagnostic performance when models receive patient history plus imaging findings [10, 11]. Pediatric neuroradiology is particularly challenging due to age-dependent prevalence, rare congenital/metabolic entities, and overlapping phenotypes [1, 2]. The relatively strong performance of leading models in this constrained setting suggests that when imaging findings are clearly articulated in report-like language, LLMs may leverage internal medical knowledge and pattern constraints to propose plausible diagnoses and differentials.

Table 2

Diagnostic performance of radiologists and large language models on 100 fictional paediatric neuroradiology cases

		Min-Max	Median(75P-25P)	Mean \pm SD
Pediatric radiologist	False			46
	True			54
Pediatric radiologist DDxScore		1.0-5.00	4.0 (4.0-2.0)	3.1 \pm 1.4
General radiologist 1	False			53
	True			47
General radiologist 1 DDxScore		1.0-5.00	3.0 (4.0-2.0)	3.0 \pm 1.3
General radiologist 2	False			56
	True			44
General radiologist 2 DDxScore		1.0-5.00	3.0 (4.0-2.0)	3.0 \pm 1.3
GPT5.2 Thinking	False			20
	True			80
GPT5.2 Thinking DDxScore		2.0-5.00	4.0 (4.0-3.0)	3.8 \pm 0.8
GPT5.2 Instant	False			32
	True			68
GPT5.2 Instant DDxScore		2.0-5.00	4.0 (4.0-3.0)	3.7 \pm 0.9
GPT5.2 Auto	False			20
	True			80
GPT5.2 Auto DDxScore		3.0-5.00	4.0 (5.0-4.0)	4.2 \pm 0.7
Gemini 3 Thinking	False			20
	True			80
Gemini 3 Thinking DDxScore		2.0-5.00	5.0 (5.0-4.0)	4.2 \pm 0.9
Gemini 3 Pro	False			52
	True			48
Gemini 3 Pro DDxScore		1.0-5.00	3.0 (5.0-3.0)	3.5 \pm 1.1
Claude 4.5 Opus Thinking	False			24
	True			76
Claude 4.5 Opus Thinking DDxScore		2.0-5.00	5.0 (5.0-4.0)	4.3 \pm 0.9
Claude 4.5 Opus	False			28
	True			72
Claude 4.5 Opus DDxScore		3.0-5.00	4.0 (5.0-3.0)	4.2 \pm 0.8

DDxScore: Differential Diagnoses Score, SD: Standart Deviation, P: Percentile

Compared with JUNG et al. [14] who evaluated LLMs on pediatric radiology cases derived from textbook-style descriptions and examined the effect of adding clinical presentation, our study is narrower (pediatric neuroradiology), uses report-like structured text authored to mirror clinical reporting, includes seven current model variants and three radiologist readers, and applies a clinically oriented ordinal DDxScore capturing partial utility beyond top-1 correctness. Likewise, relative to ABDUL SAMI et al. [15] who assessed text-based pediatric radiology questions using binary scoring, we evaluated a more practice-proximate task (open-ended diagnosis generation from structured reports) and included both human comparators and graded differential quality.

The DDxScore findings are clinically meaningful. Several models frequently produced differentials that were structured and plausible, sometimes listing the correct diagnosis even when not chosen as the top answer. In real practice, prompting consideration of high-stakes alternatives can be valuable for education, triage, and decision support, provided use remains audited and human-supervised [9]. Mode-dependent performance – higher accuracy and DDxScore in thinking/auto variants – supports the concept that test-time reasoning structure influences radiology performance [19]. WIND et al. [19] demonstrated that multi-step retrieval and reasoning can improve radiology question answering; our findings similarly suggest that deeper deliberation is beneficial for complex diagnostic synthesis in pediatric neuroradiology.

We would also like to acknowledge the limitations of this study. Firstly, while fictional structured text-only cases enhance standardization, they may overestimate performance when compared to real-world variability, incomplete histories, ambiguity, and heterogeneous reporting. Second, training-data opacity prevents exclusion of distributional overlap, which could inflate performance. Third, the text-only format omits image interpretation and thus does not generalize to image-based workflows; radiologists may be disadvantaged without visual pattern recognition. Fourth, results may be sensitive to prompt phrasing, web-interface settings, and model updates. Fifth, DDxScore includes subjective judgment and was scored by a limited evaluator set; independent scoring and interrater reliability would strengthen robustness.

Future studies should replicate these findings on real-world pediatric neuro-radiology cases including images, ideally multicentre and prospectively designed. Pre-registration of prompts, model versions, and evaluation protocols would improve reproducibility. Future evaluations should assess calibration (uncertainty vs. correctness), interrater reliability of scoring, and human-in-the-loop outcomes – i.e., whether LLM assistance improves clinician performance without unsafe over-reliance.

In conclusion, several LLM variants matched or exceeded radiologists in top-1 diagnosis and differential quality on structured, text-only pediatric neuroradiology cases while responding substantially faster. Performance was configuration-

dependent, with thinking/auto modes outperforming instant variants, supporting cautious exploration for education and audited decision support pending real-world image-based validation.

REFERENCES

- [1] ROSSI A., M. ARGYROPOULOU, D. ZLATAREVA et al. (2023) European recommendations on practices in pediatric neuroradiology: consensus document from the European Society of Neuroradiology (ESNR), European Society of Paediatric Radiology (ESPR) and European Union of Medical Specialists Division of Neuroradiology (UEMS), *Pediatr. Radiol.*, **53**(1), 159–168.
- [2] BHATIA A., F. KHALVATI, B. B. ERTL-WAGNER (2024) Artificial intelligence in the future landscape of pediatric neuroradiology: opportunities and challenges, *AJNR Am. J. Neuroradiol.*, **45**(5), 549–553.
- [3] THIRUNAVUKARASU A. J., D. S. J. TING, K. ELANGOVAN et al. (2023) Large language models in medicine, *Nat. Med.*, **29**(8), 1930–1940.
- [4] GÜNEŞ Y. C., T. CESUR, E. ÇAMUR et al. (2025) Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition, *Diagn. Interv. Radiol.*, **31**(2), 111–129.
- [5] GÜNEŞ Y. C., T. CESUR (2024) Diagnostic accuracy of large language models in the European board of interventional radiology examination (EBIR) sample questions, *Cardiovasc. Intervent. Radiol.*, **47**(6), 836–837.
- [6] ÇAMUR E., T. CESUR, Y. C. GÜNEŞ (2024) Can large language models be new supportive tools in coronary computed tomography angiography reporting? *Clin. Imaging*, **114**, 110271.
- [7] BHAYANA R., S. KRISHNA, R. R. BLEAKNEY (2023) Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations, *Radiology*, **307**(5), e230582.
- [8] SUTHAR P. P., A. KOUNSAL, L. CHHETRI et al. (2023) Artificial Intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR) "Case of the Month", *Cureus*, **15**(8), e43958.
- [9] AKINCI D'ANTONOLI T., A. STANZIONE, C. BLUETHGEN et al. (2024) Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions, *Diagn. Interv. Radiol.*, **30**(2), 80–90.
- [10] HORIUCHI D., H. TATEKAWA, T. SHIMONO et al. (2024) Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases, *Neuroradiology*, **66**(1), 73–79.
- [11] UEDA D., Y. MITSUYAMA, H. TAKITA et al. (2023) ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes, *Radiology*, **308**(1), e231040.
- [12] GUNES Y. C., T. CESUR (2025) The diagnostic performance of large language models and general radiologists in thoracic radiology cases: a comparative study, *J. Thorac. Imaging*, **40**(3), e0805.
- [13] CESUR T., Y. C. GUNES, E. CAMUR et al. (2025) Empowering radiologists with

- ChatGPT-4o: comparative evaluation of large language models and radiologists in cardiac cases, *J. Thorac. Imaging*, **40**(6), e0846.
- [14] JUNG J., M. PHILLIPI, B. TRAN et al. (2025) Accuracy of large language models in generating differential diagnosis from clinical presentation and imaging findings in pediatric cases, *Pediatr. Radiol.*, **55**(9), 1927–1933.
- [15] ABDUL SAMI M., M. ABDUL SAMAD, K. PAREKH et al. (2024) Comparative accuracy of ChatGPT 4.0 and Google Gemini in answering pediatric radiology text-based questions, *Cureus*, **16**(10), e70897.
- [16] ÇAMUR E., T. CESUR, Y. C. GÜNEŞ et al. (2026) The performance of large language models in bone tumour imaging: comparative analysis with radiologists using text and image-based evaluation, *C. R. Acad. Bulg. Sci.*, **79**(1), 95–102.
- [17] BOSSUYT P. M., J. B. REITSMA, D. E. BRUNS et al. (2015) STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies, *Clin. Chem.*, **61**(12), 1446–1452.
- [18] CESUR T., Y. C. GÜNEŞ (2024) Optimizing diagnostic performance of ChatGPT: the impact of prompt engineering on thoracic radiology cases, *Cureus*, **16**(5), e60009.
- [19] WIND S., J. SOPA, D. TRUHN et al. (2025) Multi-step retrieval and reasoning improve radiology question answering with large language models, *NPJ Digit. Med.*, **8**(1), 790.

¹*Department of Radiology, Mamak State Hospital, Üreğil 06270, Ankara, Türkiye*
e-mail: turaycesur93@gmail.com, <https://orcid.org/0000-0002-2726-8045>

²*Department of Radiology, Kirikkale Yuksek Ihtisas Hospital, Baglarbasi 71300, Kirikkale, Türkiye*
e-mail: gunesyasincelal@gmail.com, <https://orcid.org/0000-0001-7631-854X>

³*Department of Radiology, 29 Mayıs State Hospital, Dikmen 06460, Ankara, Türkiye*
e-mail: eren.camur@outlook.com, <https://orcid.org/0000-0002-8774-5800>

⁴*Department of Pediatrics, Thracian University, Merkez 22030, Edirne, Türkiye*
e-mail: drgulaycesur@gmail.com, <https://orcid.org/0000-0003-4496-1228>

⁵*Department of Pediatric Radiology, Aydin Adnan Menderes University, Efeler 09100, Aydin, Türkiye*
e-mails: gtuzcu@adu.edu.tr, <https://orcid.org/0000-0002-3957-1770>
avni.merter.keceli@adu.edu.tr, <https://orcid.org/0000-0002-9412-6733>