Original Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2026 (6 Jul);40.71. https://doi.org/10.47176/mjiri.40.71

Interdisciplinary Evaluation of Urogenital Radiological Findings: Human Expert vs Multimodal Generative AI Classification Performance

Büşra Asena Torun[1]*, Alpkon Torun[2], Mustafa Serdar Çağlayan[2]

Received: 27 May 2026 Revised: 19 Jun 2026 Accepted: 29 Jun 2026 Published: 6 Jul 2026

Abstract

Background: Multimodal generative artificial intelligence (AI) systems are increasingly evaluated for medical image classification, but their performance in interdisciplinary urogenital radiological assessment remains uncertain. This study aimed to compare the overall and domain-specific classification performance and uncertainty patterns of two domain-specific human specialists and three multimodal generative AI systems in a standardized single-image urogenital radiology task.

Methods: This exploratory pilot comparative study evaluated 60 publicly available anonymized radiological cases selected from Radiopaedia, including 20 urological anomaly/variation cases, 20 gynecological anomaly/variation cases, and 20 normal anatomy cases. Two human specialists and three multimodal AI systems (GPT-4o, Gemini 1.5 Pro, and Microsoft Copilot) independently classified randomized single images into predefined categories under blinded conditions. Overall accuracy was summarized with 95% confidence intervals, while domain-specific accuracy and uncertain/unable-to-classify responses were analyzed descriptively.

Results: Overall accuracy was highest for Gemini 1.5 Pro (36/60, 60.0%; 95% CI, 47.4-71.4), followed by the urologist (34/60, 56.7%; 95% CI, 44.1-68.4), GPT-4o (31/60, 51.7%; 95% CI, 39.3-63.8), the gynecologic oncologist (29/60, 48.3%; 95% CI, 36.2-60.7), and Copilot (19/60, 31.7%; 95% CI, 21.3-44.2). Human specialists performed best within their own domains. Gemini performed strongly in urological and gynecological categories but had low normal-category accuracy (2/20, 10.0%), suggesting overclassification.

Conclusion: In this standardized single-image pilot task, multimodal AI systems showed potential as supportive classification tools but also demonstrated clinically relevant limitations, particularly false-positive interpretation of normal anatomy. These findings should be interpreted as hypothesis-generating and not as evidence of autonomous diagnostic capability.

Keywords: Artificial Intelligence, Decision Support Systems, Clinical, Human-AI Interaction, Medical Image Interpretation, Multimodal Large Language Models

Cite this article as: Torun BA, Torun A, Çağlayan MS. Interdisciplinary Evaluation of Urogenital Radiological Findings: Human Expert vs Multimodal Generative AI Classification Performance. *Med J Islam Repub Iran.* 2026 (6 Jul);40:71. https://doi.org/10.47176/mjiri.40.71

Introduction

Recent advances in multimodal generative artificial intelligence (AI) have enabled systems capable of simultaneously processing visual and textual information, expanding their applicability beyond traditional prediction-based models into complex interpretation tasks. These developments have introduced new opportunities for classification problems involving heterogeneous and interdisciplinary datasets, particularly in domains where structural complexity and contextual variability are prominent (1, 2).

Medical imaging represents one such domain, where image interpretation often requires not only pattern recognition but also domain-specific anatomical knowledge (3). While conventional AI systems have primarily focused on single-domain classification tasks, multimodal generative AI models offer the potential to integrate knowledge across different anatomical and clinical contexts (4, 5).

Corresponding author: Dr Büşra Asena Torun, b.asena.torun@gmail.com

[1.] Department of Gynecologic Oncology, Adana City Training and Research Hospital, University of Health Sciences, Adana, Turkey
[2.] Department of Urology, Faculty of Medicine, Hitit University, Çorum, Turkey

Key Messages:

↑What is "already known" in this topic:

Multimodal large language models are increasingly evaluated for medical image interpretation, but their reliability in urogenital radiological assessment and cross-disciplinary clinical contexts remains uncertain.

→What this article adds:

This exploratory pilot study compares human specialists and multiple multimodal AI models in a standardized single-image urogenital radiology task, highlighting both the supportive potential and current limitations of AI-assisted assessment.

This capability may be particularly relevant in anatomically intertwined systems such as the urogenital region, where radiological findings frequently involve structures that fall within the expertise of multiple clinical disciplines.

Despite the growing interest in the application of generative AI in radiological workflows, existing studies have largely evaluated performance within isolated clinical domains or focused on specific disease detection tasks (6, 7). Comparative investigations that examine the classification behavior of generative AI systems across interdisciplinary anatomical categories, particularly in direct comparison with domain-specific human experts, remain limited.

Furthermore, the decision-making characteristics of generative AI systems—such as sensitivity to anomaly detection, confidence calibration, and classification tendencies—have not been fully characterized in anatomically complex settings. Understanding these behavioral patterns is essential for assessing the feasibility of integrating AI-based outputs into interdisciplinary decision-support environments.

In this context, the present study aims to evaluate the classification performance of multimodal generative AI systems in urogenital radiological image interpretation through a blinded comparison with domain-specific human experts. By analyzing domain-specific accuracy, cross-domain performance, and classification uncertainty, this study seeks to provide insight into the potential role and limitations of generative AI as a complementary tool in anatomically complex classification tasks.

The evolution of artificial intelligence in radiology has transitioned from early rule-based computer-aided detection (CAD) systems to advanced deep learning architectures (8). While traditional convolutional neural networks (CNNs) achieved high performance in single-domain tasks (9), the recent emergence of multimodal generative AI marks a historical shift toward systems capable of integrating cross-disciplinary anatomical knowledge (10).

Recent developments in artificial intelligence have significantly impacted medical image analysis, particularly in classification and detection tasks (11). Deep learning-based models have demonstrated performance comparable to human experts in various imaging domains, including dermatology (12), and oncology (13), highlighting the potential of AI systems in visual diagnostic support tasks.

In radiology, artificial intelligence has increasingly been integrated into workflows for image interpretation, anomaly detection, and report generation. Previous studies have demonstrated the ability of AI systems to assist in diagnostic processes by improving detection sensitivity and reducing interpretive variability. More recently, the emergence of multimodal generative AI models capable of integrating visual and textual information has further expanded these capabilities beyond traditional classification frameworks.

Despite these advancements, most existing studies have evaluated AI performance within single clinical domains or focused on disease-specific detection tasks. Investigations exploring cross-domain anatomical classification—particularly in anatomically complex systems involving multiple clinical specialties—remain limited (6, 7).

Comparative analyses between human experts and AI systems have also highlighted key differences in decision-making patterns. While AI systems may demonstrate high sensitivity in anomaly detection, challenges related to specificity, confidence calibration (14), and generalization across unfamiliar domains have been reported (15, 16). These limitations underscore the importance of understanding AI behavior in interdisciplinary contexts.

Accordingly, there remains a need for studies that evaluate the performance of multimodal generative AI systems in anatomically intertwined regions requiring expertise from multiple clinical disciplines. The present study seeks to address this gap by directly comparing AI-based classification with domain-specific human expertise in the interpretation of urogenital radiological findings.

Methods

This study was designed as an exploratory pilot comparative evaluation framework to assess the classification performance of multimodal generative artificial intelligence systems against domain-specific human experts in the interpretation of urogenital radiological images.

A total of 60 radiological cases were selected from Radiopaedia, a publicly accessible educational imaging database. The dataset included images from multiple modalities, including computed tomography (CT), magnetic resonance imaging (MRI), ultrasonography, X-ray, and scintigraphy.

Cases were equally distributed across three predefined anatomical categories:

- Urological anomaly or variation
- Gynecological anomaly or variation
- Normal anatomy

Each category consisted of 20 cases to ensure balanced representation across domains.

To minimize interpretive bias, only single representative cross-sectional images were used for each case. Images containing diagnostic annotations, arrows, or textual hints were excluded.

Each image was presented as an independent classification task. Evaluators were required to assign the image to one of the following predefined categories:

- Urological anomaly/variation
- Gynecological anomaly/variation
- Normal anatomy
- Uncertain / unable to classify

The same standardized prompt and the same response options were used for all AI systems. The “uncertain / unable to classify” option was included as a predefined fourth response category to avoid forcing evaluators to assign a normal or abnormal label when the isolated image did not provide sufficient confidence. This category allowed assessment of whether human and AI evaluators would recognize uncertainty rather than overclassify normal or ambiguous findings.

[DOI: 10.47176/mjiri.40.71]

Reference classifications were determined based on the original case descriptions provided in the source database.

Two domain-specific human experts participated in the evaluation:

- One urology specialist
- One gynecologic oncology specialist

Experts were blinded to the case distribution and reference classifications.

Three multimodal generative AI systems were evaluated:

- GPT-4o (OpenAI)
- Gemini 1.5 Pro (Google)
- Copilot (Microsoft)

All AI evaluations were performed using their multimodal versions available as of January 2026. When exact backend version identifiers were not publicly available through the web interfaces, the platform name, model designation, and access period were reported. Therefore, January 2026 should be interpreted as the access period of evaluation rather than as a standalone model version identifier.

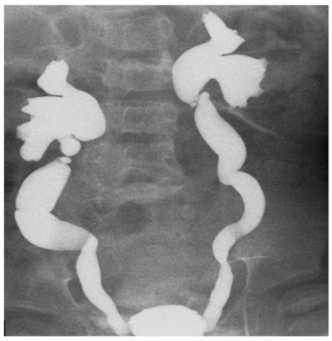
All cases were presented in randomized order under blinded conditions. Human experts and AI systems independently performed the classification tasks without access to prior responses or contextual clinical information (Figures 1 and 2).

Classification performance was evaluated using:

- Overall accuracy
- Domain-specific accuracy
- Uncertainty response rate
- Inter-rater agreement (Cohen's kappa)

Categorical comparisons were analyzed using chi-square or Fisher's exact tests when appropriate. Within-evaluator domain comparisons were performed using McNemar's test. Statistical significance was defined as $P<0.05$.

Overall accuracy was additionally reported as counts and percentages with Wilson 95% confidence intervals.



Which anatomical condition is demonstrated in this image?

Response options:

A) Urological anomaly/variation
B) Gynecological anomaly/variation
C) Normal anatomy
D) Uncertain / unable to classify

Figure 1. Example of the evaluation framework used for image classification tasks.

Which anatomical condition is demonstrated in this image'

Response options:

A) Urological anomaly/variation
B) Gynecological anomaly/variation
C) Normal anatomy
D) Uncertain / unable to classify

Figure 2. Example of the evaluation framework used for image classification tasks.

Results

Overall classification performance varied across evaluators, demonstrating heterogeneous behavior among both human experts and AI systems. Gemini achieved the highest overall accuracy (60.0%), followed by the urologist (56.7%), GPT-4o (51.7%), the gynecologic oncologist (48.3%), and Copilot (31.7%) (Figure 3).

Domain-specific analysis revealed that human experts demonstrated superior performance within their respective fields. The urologist achieved an accuracy of 80% in urological cases ($P=0.004$), while the gynecologic oncologist demonstrated an accuracy of 75% in gynecological cases ($P=0.018$). However, performance differences became more pronounced in cross-domain evaluations. Human experts showed reduced accuracy outside their primary areas of specialization, with performance levels approaching those of certain AI systems (Figures 4a and 4b).

Among AI systems, Gemini exhibited strong domain-level performance, achieving 80% accuracy in urological cases and 90% accuracy in gynecological cases. In gynecological classifications, Gemini surpassed the gynecologic oncologist, while demonstrating parity with the urologist in urological scenarios.

In the classification of normal anatomy, AI systems demonstrated notable limitations. Gemini achieved an accuracy rate of only 10% in normal cases ($P=0.006$), indicating a strong tendency toward false-positive classification of anatomical variations (Figure 4c).

The Copilot model exhibited a high rate of uncertainty responses, selecting the "uncertain" category in 40% of cases ($P<0.001$), significantly higher than both human evaluators and other AI systems.

Inter-rater agreement analysis demonstrated moderate concordance across evaluators, with variability reflecting

[DOI: 10.47176/mjiri.40.71]

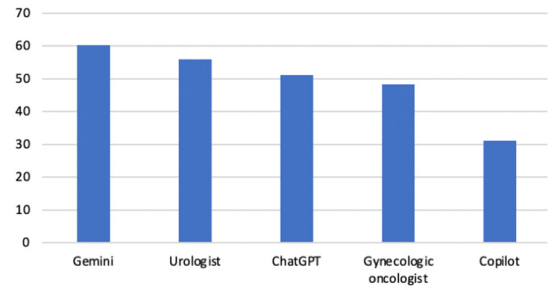


Figure 3. Overall classification accuracy (%) across evaluators.

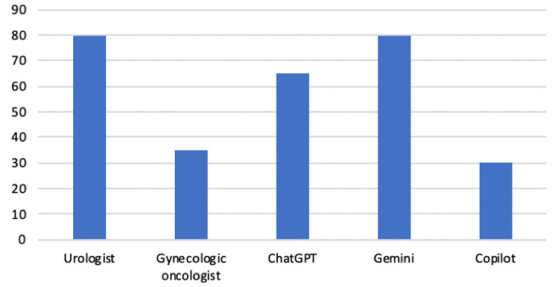


Figure 4A. Accuracy (%) in urological cases.

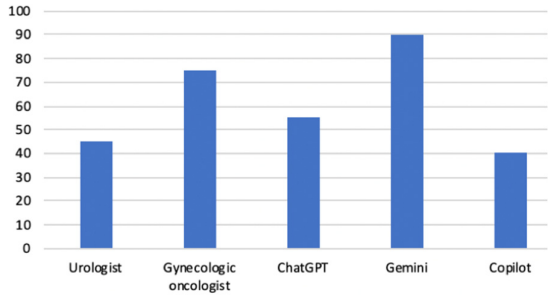


Figure 4B. Accuracy (%) in gynecological cases.

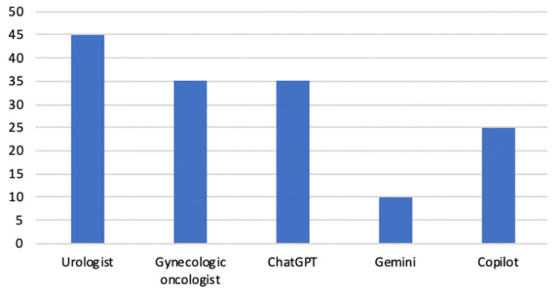


Figure 4C. Accuracy (%) in normal anatomy.

domain specialization effects and differing classification strategies (Table 1).

Table 1. Accuracy metrics across human specialists and AI models

Evaluator	Overall	Uro	Gyn	Normal
Urologist	56.7	80.0	45.0	45.0
Gyn. Oncologist	48.3	35.0	75.0	35.0
GPT-4o	51.7	65.0	55.0	35.0
Gemini 1.5 Pro	60.0	80.0	90.0	10.0
Copilot	31.7	30.0	40.0	25.0

Note: Uro: Urological, Gyn: Gynecological.

The findings of this study reveal distinct behavioral differences between domain-specific human expertise and multimodal generative AI systems in anatomical classification tasks.

Human experts demonstrated strong performance within their respective domains, reflecting the impact of specialized knowledge in structured anatomical interpretation. However, their accuracy declined in out-of-domain evaluations, suggesting limitations in cross-domain generalization.

The relatively low Normal-category accuracy of the urologist should not be interpreted as insufficient urological expertise. In this interdisciplinary task, the urologist was not evaluating only urological pathology, but was asked to classify urological findings, gynecological findings, and normal anatomy. Therefore, overall accuracy alone does not directly represent urological domain expertise. The urologist achieved 80% accuracy in urological cases, indicating preserved domain-specific performance despite lower accuracy in normal or non-urological categories. In contrast, AI systems exhibited a more uniform but heterogeneous performance profile across anatomical categories. Gemini demonstrated high sensitivity in detecting anomalies across both domains, achieving domain-level parity with human experts in urological cases and surpassing the gynecologic oncologist in gynecological classifications.

However, this sensitivity appeared to be associated with a trade-off in specificity. The notably low accuracy of AI systems—particularly Gemini—in identifying normal anatomy indicates a pronounced false-positive tendency. This behavior reflects an overclassification bias, where structurally normal findings are interpreted as anomalous.

Such classification tendencies may arise from the probabilistic decision-making mechanisms inherent in generative AI systems, which prioritize anomaly detection over conservative interpretation in the absence of contextual constraints (17, 18).

Additionally, the high uncertainty response rate observed in Copilot suggests an alternative decision profile characterized by cautious classification under ambiguity. While this may reduce false-positive interpretations, it also limits practical utility in decision-support scenarios.

Overall, these findings highlight a fundamental distinction between human and AI interpretive strategies: domain specialization versus generalized pattern sensitivity (19). While human experts rely on contextual anatomical familiarity, AI systems leverage broad training distributions that enable cross-domain recognition but may compromise specificity (20, 21).

Understanding these behavioral patterns is essential for evaluating the feasibility of integrating AI-based outputs into interdisciplinary classification workflows.

Discussion

The observed performance patterns suggest that multimodal generative AI systems may hold value as supportive tools in interdisciplinary anatomical interpretation rather than as autonomous decision-makers. AI systems demonstrated the ability to recognize patterns across mul-

[DOI: 10.47176/mjiri.40.71]

tiple anatomical domains, occasionally achieving performance levels comparable to or exceeding those of domain-specific experts in focused classification tasks. This indicates potential utility in enhancing interdisciplinary awareness, particularly in anatomically complex regions where findings may extend beyond the primary expertise of a single specialty.

However, the pronounced tendency toward overclassification, particularly in the misinterpretation of normal anatomy, highlights the risks associated with unsupervised deployment. False-positive outputs in clinical environments may lead to unnecessary diagnostic procedures, increased workload, or misdirected clinical attention. In this context, the most appropriate integration strategy for generative AI lies within a complementary decision-support framework. AI-generated classifications may function as a secondary interpretive layer, drawing attention to potential anomalies or cross-domain considerations that might otherwise be overlooked. Such an approach aligns with human-centered AI integration models, in which automated outputs augment—but do not replace—expert clinical judgment (21, 22). Accordingly, the role of multimodal generative AI in urogenital radiological interpretation should be conceptualized as supportive rather than substitutive, operating under continuous human oversight.

Several limitations of this study should be acknowledged. First, the relatively small sample size may limit the generalizability of the findings. Although balanced case selection allowed for comparative evaluation across domains, larger datasets would provide more robust estimates of performance differences. Second, the use of publicly available educational imaging cases may not fully reflect the complexity and variability encountered in real-world clinical practice. Third, the evaluation was based on single representative cross-sectional images rather than full imaging series. While this approach enabled standardized comparison across evaluators, it differs from routine clinical workflows that involve multi-slice and dynamic interpretation. As of June 2026, general-purpose MLLMs are not yet clinically established or validated for comprehensive DICOM-based interpretation of full CT or MRI series in a multi-slice, multiplanar, and clinically contextualized manner. Therefore, the present design should be interpreted as a standardized single-image classification and pattern-recognition task rather than a simulation of complete radiological diagnostic workflow.

Additionally, the performance of generative AI systems may vary depending on model versions, temporal updates, and model-specific hallucination tendencies (23-25). As such, the findings of this study should be interpreted as time-specific.

Finally, the study focused on anatomical classification rather than comprehensive diagnostic decision-making. Therefore, the results should be interpreted as indicative of interpretive behavior rather than clinical diagnostic capability.

Conclusion

This study demonstrates that multimodal generative artificial intelligence systems exhibit meaningful potential in the classification of urogenital radiological findings across interdisciplinary anatomical domains. While AI systems showed the ability to achieve domain-level performance comparable to human experts in certain classification tasks, their limitations in identifying normal anatomy revealed a pronounced overclassification tendency. This finding underscores the importance of maintaining human oversight when integrating AI into interpretive workflows. Human experts demonstrated strong domain-specific performance but showed reduced accuracy in out-of-domain evaluations, highlighting the complementary strengths of specialized expertise and generalized pattern recognition.

Taken together, these results suggest that the optimal role of multimodal generative AI lies not in autonomous decision-making, but in functioning as a supportive interpretive layer within interdisciplinary classification environments. Future research involving larger datasets and real-world imaging workflows may further clarify the potential of AI-assisted interpretation in anatomically complex systems.

Acknowledgment

None.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contributions

Büşra Asena Torun contributed to the study conception and design, data collection, image selection, evaluation of radiological cases, interpretation of results, and drafting of the manuscript. Mustafa Serdar Çağlayan contributed to study design, clinical evaluation, interpretation of findings, critical revision of the manuscript, and supervision. Alpkon Torun contributed to data organization, statistical analysis, interpretation of results, manuscript editing, and critical review. All authors read and approved the final version of the manuscript.

Ethical Considerations

This study used publicly available anonymized educational radiological cases and did not involve direct patient contact, intervention, animal experimentation, or identifiable patient data. Therefore, institutional ethical approval and informed consent were not required.

Funding Support

This research received no external funding. The article processing charge, if applicable, was funded by the authors.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable re-

[DOI: 10.47176/mjiri.40.71]

quest. The radiological cases used in the study were obtained from publicly available educational sources, in accordance with the relevant source conditions.

AI Use Statement

Multimodal generative artificial intelligence systems were evaluated as the subject of this study, as described in the Methods section. AI-assisted tools were not used for data analysis or clinical interpretation. During manuscript preparation, AI-assisted language editing support was used only to improve clarity and grammar. All scientific content, results, interpretations, and references were critically reviewed and approved by the human authors, who take full responsibility for the final manuscript

References

1. OpenAI. GPT-4 technical report. Preprint. arXiv. Posted March 15, 2023. doi:10.48550/arXiv.2303.08774.
2. Gemini Team. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint. arXiv. Posted March 8, 2024. doi:10.48550/arXiv.2403.05530.
3. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE Inst Electr Electron Eng. 2021;109(5):820-838. doi:10.1109/JPROC.2021.3054390.
4. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0.
5. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-265. doi:10.1038/s41586-023-05881-4.
6. Jung HK, Kim K, Park JE, Kim N. Image-based generative artificial intelligence in radiology: comprehensive updates. Korean J Radiol. 2024;25(11):959-981. doi:10.3348/kjr.2024.0392.
7. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. 2024;26:e59505. doi:10.2196/59505.
8. Avanzo M, Stancanello J, Pirrone G, Drigo A, Retico A. The evolution of artificial intelligence in medical imaging: from computer science to machine and deep learning. Cancers (Basel). 2024;16(21):3702. doi:10.3390/cancers16213702.
9. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500-510. doi:10.1038/s41568-018-0016-5.
10. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents. Preprint. arXiv. Posted April 13, 2022. doi:10.48550/arXiv.2204.06125.
11. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health. 2023;11:1273253. doi:10.3389/fpubh.2023.1273253.
12. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29(8):1836-1842. doi:10.1093/annonc/mdy166.
13. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging. 2020;39(4):1184-1194. doi:10.1109/TMI.2019.2945514.
14. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6-11; Sydney, Australia. Proceedings of Machine Learning Research; 2017. p. 1321-1330.
15. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2.
16. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-2182. doi:10.1038/s41591-021-01595-0.
17. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):248. doi:10.1145/3571730.
18. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. Preprint. arXiv. Posted November 9, 2023. doi:10.48550/arXiv.2311.05232.
19. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 2021;4:31. doi:10.1038/s41746-021-00385-9.
20. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intell. 2020;2(11):665-673. doi:10.1038/s42256-020-00257-z.
21. Shneiderman B. Human-Centered AI. Oxford University Press; 2022. doi:10.1093/oso/9780192845290.001.0001.
22. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7.
23. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? Preprint. arXiv. Posted July 18, 2023. doi:10.48550/arXiv.2307.09009.
24. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol. 2023;96(1150):20220878. doi:10.1259/bjr.20220878.
25. von der Stück MS, Vuskov R, Westfechtel S, Siepmann R, Kuhl C, Truhn D, et al. Visual large language models in radiology: a systematic multimodel evaluation of diagnostic accuracy and hallucinations. Life (Basel). 2026;16(1):66. doi:10.3390/life16010066.

[DOI: 10.47176/mjiri.40.71]