Research Article ISSN 2835-3587

Radiology & Imaging Journal

Reliability of ChatGPT in the Evaluation of Voiding Cystourethrograms: Comparison with Experts in Cases of Bulbar Stricture and Normal Studies

Magnum Adriel Santos Pereira*, João Jorge Saab, Marina Grzybowski Paranhos, Raul Loures, Daniel Charret Diegues, Lorella Miranda Auricchio, Luis Augusto Seabra Rios, and Wagner Aparecido França

Urology Department, Hospital do Servidor Público Estadual (HSPE/IAMSPE), São Paulo, Brazil.

***Correspondence:**
Magnum Adriel Santos Pereira, MD, Rua Pedro de Toledo, 1800 – Vila Clementino – São Paulo, SP – CEP 04039-000 – Brazil, Telephone: +55 092991231854, Phone: +55 (11) 4573-8000.

Received: 08 Sep 2025; **Accepted:** 15 Oct 2025; **Published:** 24 Oct 2025

Citation: Magnum Adriel Santos Pereira, João Jorge Saab, Marina Grzybowski Paranhos, et al. Reliability of ChatGPT in the Evaluation of Voiding Cystourethrograms: Comparison with Experts in Cases of Bulbar Stricture and Normal Studies. Radiol Imaging J. 2025; 4(4); 1-6.

ABSTRACT

Background: *The use of large language models (LLMs) for medical image interpretation has expanded rapidly, yet clinical validation remains limited. We evaluated ChatGPT's performance in interpreting voiding cystourethrograms (VCUGs) for bulbar urethral stricture.*

Objective: *To assess the diagnostic accuracy and treatment recommendations generated by ChatGPT when interpreting VCUG images, compared with reconstructive urology experts and with the procedure actually performed.*

Methods: *We conducted a retrospective cross-sectional study at a tertiary public hospital. A total of 51 VCUGs were analyzed: 41 confirmed bulbar strictures and 10 normal studies. De-identified, representative static frames from retrograde and voiding phases were presented to ChatGPT (version 4.0 – 1.2025.105) in independent chats using a standardized English prompt. Two reconstructive urologists (GURS members; >50 urethral surgeries/year) independently reviewed all cases. Performance metrics included sensitivity, specificity, accuracy, predictive values, and Cohen's kappa for agreement.*

Results: *ChatGPT correctly identified 40/41 bulbar strictures (sensitivity 97.56%) but labeled all 10/10 normal VCUGs as strictures (specificity 0%). Overall accuracy was 78.43%, positive predictive value 80%, negative predictive value 0%, and Cohen's kappa 0.51 (moderate agreement). ChatGPT tended to overcall strictures, limiting its usefulness for triage when normal studies are prevalent. When the anatomic location was correctly identified, suggested treatments were generally concordant with contemporary guideline-based management.*

Conclusion: *ChatGPT showed very high sensitivity but null specificity for bulbar stricture detection on VCUG static frames, indicating substantial limitations for independent diagnostic use. The model may serve as a supervised aid where specialist access is scarce, while future multimodal models specifically trained on urologic imaging may achieve better balance between sensitivity and specificity.*

Keywords
Artificial intelligence, Bulbar urethral stricture, ChatGPT, Diagnostic accuracy, VCUG.

Introduction
Bulbar stricture refers to narrowing of the bulbar portion of the urethra, which is part of the male anterior urethra [1,2]. Among anterior urethral strictures, the bulbar location is the most frequent, accounting for about 47% of cases in large series of adult patients [2,3]. Symptoms may be confused with other urological conditions, hindering the initial diagnosis [4]. Difficulty urinating, a weak urinary stream, the sensation of incomplete bladder emptying, increased urinary frequency, and the need to strain to urinate are common [1,5.] Untreated cases may progress to recurrent urinary

tract infections, urinary retention, detrusor failure, and even renal impairment [6,7]. The disease causes marked obstructive urinary symptoms, potentially leading to complications and affecting patients' quality of life and emotional health.

In younger patients, trauma and idiopathic causes predominate, whereas in the elderly, iatrogenic causes are more common [8,9]. In developed countries, iatrogenesis is the main cause; in developing countries, trauma and infections still play a significant role [9,10].

Retrograde urethrography (RUG) is performed by the retrograde introduction of contrast through the distal urethra, followed by radiographic images to visualize the path of the contrast along the anterior and posterior urethra [11], whereas voiding cystourethrography (VCUG) is performed by introducing contrast medium into the bladder, usually via a urinary catheter, followed by fluoroscopic images during voiding to assess the path of contrast through the urethra [11,12].

RUG is the standard method for evaluating traumatic and inflammatory lesions and strictures of the male urethra, allowing the location, extent, and multifocality of strictures to be determined, as well as assisting in surgical planning [11,13]. When combined, the two techniques provide complete assessment of the urethra, especially in complex cases or posterior strictures [13]. Accurate stricture assessment is essential for choosing between endoscopic treatments (indicated for short strictures) and open surgeries (required for long or recurrent strictures) [14,15].

Interpretation of studies by an experienced radiologist or urologist is crucial to differentiate normal findings from pathological changes, avoiding misdiagnoses and inappropriate treatments [12,13]. Specialists use imaging findings to define the best surgical approach, assess the need for additional tests, and monitor treatment success, especially after procedures such as urethroplasty [16,17].

Artificial intelligence (AI), especially deep learning techniques such as convolutional neural networks, has already achieved performance comparable to human experts in tasks such as diagnosis, segmentation, and classification of medical images [18,19]. AI systems assist in reading histopathological images, reduce inter-physician variability, and increase disease detection rates, in addition to supporting more precise therapeutic decisions [20,21]. Despite AI's success in static images, there has been little application in dynamic studies such as VCUG, which evaluates lower urinary tract function during voiding. AI could standardize interpretations, detect subtle anomalies, and reduce inter-rater variability [22,23]. Language models such as GPT-4 show superior performance on theoretical urology exams, but still have limitations in clinical reasoning and answer accuracy, requiring rigorous validation before practical use [24,25].

The methodological decision to restrict the study to bulbar urethral stricture was based on three main factors: (1) it is the most prevalent location of urethral strictures in adults, representing the majority of surgically treated cases; (2) the anatomy and interpretation of VCUG in the bulbar topography are more standardized and less subject to technical variations compared to penile, post-hypospadias, or complex strictures; and (3) treatment of bulbar stricture has well-established therapeutic algorithms in international guidelines, which favors comparison with plans suggested by artificial intelligence.

Therefore, the present study aims to evaluate the diagnostic accuracy and therapeutic proposal provided by the AI model ChatGPT in interpreting VCUGs of patients with bulbar stricture. The analysis will be carried out using real images, comparing them with the assessment of experts in urethral reconstruction and with the surgery actually performed, seeking to measure the AI's ability to provide safe, effective answers aligned with contemporary urological clinical practice. In addition, we intend to explore the tool's potential as diagnostic support, especially in contexts with limited access to urethral specialists, such as emergency services, primary care, or remote regions.

Methodology

This is a retrospective, cross-sectional, analytical study conducted at the Reconstructive Urology Service of the São Paulo State Public Servants' Hospital (IAMSPE). The objective of this study was to evaluate the diagnostic accuracy and therapeutic adequacy of the ChatGPT artificial intelligence model in interpreting voiding cystourethrograms (VCUGs) of patients with a confirmed diagnosis of bulbar urethral stricture.

A total of 51 VCUG exams were analyzed, 41 from patients with a confirmed diagnosis of bulbar urethral stricture and 10 exams considered normal, without urethral anatomical changes, intentionally included to evaluate not only the AI's ability to correctly detect the presence of a stricture but also its competence in recognizing normal exams, allowing calculation of specificity and simulation of a realistic clinical scenario.

Inclusion criteria were: adult male patients, exams of good technical quality, surgically confirmed diagnosis (for stricture cases), and complete documentation of the treatment performed. Normal exams were selected based on a careful review, with confirmation of the absence of anatomical changes by specialists (a urologist specialized in reconstruction and a radiologist experienced in VCUG).

Exclusion criteria were penile strictures, pan-urethral strictures, patients under 12 years of age, congenital anomalies of the genitourinary system, exams from external services, cases with multiple strictures, or exams with unsatisfactory technical quality.

Images were completely anonymized, with removal of all identifiable information. For each exam, between two static frames representative of urethral anatomy were selected, one covering the retrograde phase and another the voiding phase, choosing the best images available in each patient's archive. No image was published or disclosed.

The artificial intelligence model used was ChatGPT, version 5.0 – 1.2025.273 (18210387024). All analyses were performed on May 1, 2025. To ensure independence between cases, each VCUG was entered into a new chat window, with all memory and personalization options disabled. The objective was to simulate real use by non-specialist physicians, using only direct questions in natural language, without elaboration of complex prompts.

Interactions with the AI model were carried out in English, the language chosen because it is the main language of the scientific literature, medical guidelines, and the model's own training, ensuring better understanding of technical terms and greater consistency in responses. The standardized prompt used was as follows:
"This is an image from a voiding cystourethrogram (VCUG) of a patient with urinary complaints. Carefully analyze the image and answer the following questions: (1) Is there any visible anatomical abnormality? If yes, describe it and specify the location. (2) What is the most likely diagnosis based on the image? (3) What would be the most appropriate therapeutic approach for this specific case? (4) Justify your response based solely on the image provided. The images correspond to the same patient in different angles."

The VCUG images were blindly evaluated by two urologists specialized in reconstructive surgery, both members of GURS (Genitourinary Reconstructive Surgeons), with experience of more than 50 urethral surgeries per year. Each specialist performed his/her evaluation independently, providing the anatomical diagnosis and the therapeutic plan considered ideal for each case.

Subsequently, the surgical procedure actually performed on the patient was analyzed in order to compare the therapeutic planning suggested by the AI with the conduct performed in clinical practice.

A third specialist urologist, different from the initial evaluators, performed a final and independent analysis, assessing: (i) the correspondence between the anatomical diagnosis provided by the AI and that of the specialists; (ii) the agreement between the therapeutic plan proposed by the AI and the surgical procedure actually performed; and (iii) the AI's ability to propose plans compatible with current clinical practices and aligned with international guidelines.

Statistical analysis consisted of descriptive and inferential methods. Descriptive statistics included calculation of absolute and relative frequencies for categorical variables (such as type of diagnosis and type of plan suggested) and measures of central tendency (mean, median) and dispersion (standard deviation) for continuous variables, when applicable.

To assess the agreement between the AI diagnosis and that of specialists, as well as between the proposed plan and the surgery performed, Cohen's kappa index was used and interpreted according to the Landis and Koch scale. In addition, the following diagnostic performance metrics were calculated: sensitivity (proportion of strictures correctly identified by the AI), specificity (proportion of cases correctly classified as absence of stricture), overall accuracy, positive predictive value (PPV), and negative predictive value (NPV).

For comparisons between adequacy scores of responses, when applicable, nonparametric tests such as the Friedman test (for dependent variables) or the Kruskal–Wallis test (for independent groups) were used. The significance level adopted was 5% ($p < 0.05$), and 95% confidence intervals were calculated whenever applicable.

All data were organized in a structured spreadsheet containing the following variables: diagnosis provided by the AI, diagnosis provided by the specialists, therapeutic plan suggested by the AI, therapeutic plan proposed by the specialists, surgical procedure actually performed, and overall classification of the AI response (correct, partially correct, incorrect, or potentially dangerous). Statistical analyses were performed using Jamovi software, current version.

The project was approved by the Research Ethics Committee of IAMSPE-SP, and all participating patients signed the Informed Consent Form (ICF), in accordance with the standards of CNS Resolution 510/2016.

Results

Fifty-one voiding cystourethrography (VCUG) exams were evaluated using the ChatGPT artificial intelligence model (version 4.0 – 1.2025.105), with the objective of identifying the presence of bulbar urethral stricture and suggesting the appropriate therapeutic plan. Of these, 41 exams corresponded to cases with a confirmed diagnosis of bulbar urethral stricture, while 10 were normal exams, with absence of anatomical changes, confirmed by reconstructive specialists.

In the diagnostic analysis, the AI model's performance showed a pattern of high sensitivity, correctly identifying 40 of the 41 cases with bulbar urethral stricture (true positives), resulting in a sensitivity of 97.56%. Only one case with bulbar stricture was incorrectly classified as normal (false negative).

However, in the normal cases, the model showed significant limitations. None of the 10 normal exams were correctly recognized as such; all were erroneously interpreted as having a stricture (false positives), which resulted in a specificity of 0%. This indicates that the model failed to distinguish truly normal exams, demonstrating a tendency to overestimate the presence of bulbar stricture.

Based on the contingency matrix: True Positives (TP): 40; False Positives (FP): 10; True Negatives (TN): 0; False Negatives (FN): 1. The following metrics were calculated: Sensitivity: 97.56%; Specificity: 0%; Overall Accuracy: 78.43%; Positive Predictive Value (PPV): 80%; Negative Predictive Value (NPV): 0%; Cohen's Kappa Index: 0.51 (moderate agreement according to Landis and Koch).

These data suggest that the model has excellent ability to detect bulbar stricture when present, but a null ability to correctly rule out normal cases, which compromises its use as a broad triage tool. The tendency toward false positives increases the PPV (indicating that when there is a diagnosis, it is often true) but brings the NPV to zero, undermining confidence in negative results.

This discrepancy between sensitivity and specificity was graphically represented in histograms and summary tables and should be considered in the clinical interpretation of the model's performance.

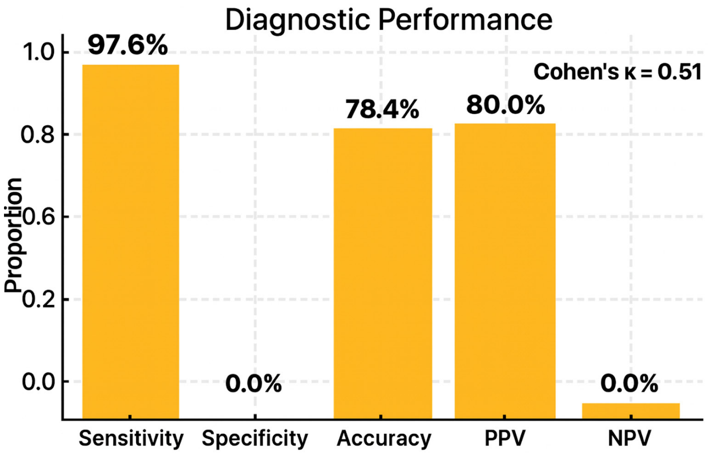

Figure 1: Bar chart illustrating ChatGPT's diagnostic performance metrics when interpreting voiding cystourethrograms (VCUG) for bulbar urethral stricture. The model achieved 97.6% sensitivity, 0.0% specificity, 78.4% overall accuracy, 80.0% positive predictive value (PPV), and 0.0% negative predictive value (NPV), with a Cohen's kappa of 0.51, indicating moderate agreement with reconstructive urology specialists. The results demonstrate high sensitivity but poor specificity, reflecting a systematic bias toward overdiagnosing strictures.

Discussion

The results of this study demonstrate that the language model based on artificial intelligence (ChatGPT) shows high sensitivity (97.56%) in detecting bulbar urethral stricture through the analysis of static images of voiding cystourethrograms (VCUG). The performance is notable, especially considering the use of natural language and the absence of complex technical prompts, simulating realistic use by non-specialist physicians.

The specificity of 0% reveals that the model completely failed to distinguish normal exams from exams with stricture. This means that any VCUG, even if normal, was interpreted by the AI as containing bulbar stricture. This bias nullifies the usefulness of the AI as a reliable diagnostic tool for screening or even for a second opinion—because it always assumes the presence of a stricture—which can lead to overdiagnoses, unnecessary anxiety, or even unwarranted surgeries.

One factor that may explain this bias is the physiological occlusion of the prostatic and membranous urethra in the retrograde phases of normal exams, which often causes confusion even for experienced physicians [26]. The AI model, by failing to integrate and compare different phases of the same exam (e.g., retrograde and voiding phases), erroneously interprets the absence of contrast in these regions as pathological obstruction. Even when informed that it is the same patient and exam, the AI does not maintain sufficient memory to correlate distinct images in an integrated manner. This demonstrates a technical limitation inherent to the model's current architecture, which does not operate with sequential visual recognition or longitudinal analysis of imaging data.

Despite the critical limitation in specificity, the Cohen's kappa index of 0.51 indicates moderate agreement between the AI and reconstructive surgery specialists. The overall accuracy of 78.43%, although impacted by the high rate of false positives, suggests room for improvement and possible future applicability, especially in scenarios with restricted access to specialized urologists. However, the model in its current form is not reliable as an independent diagnostic tool and should be used only in a supervised manner and with caution.

Regarding the therapeutic proposal, performance was considered satisfactory in cases of correct diagnosis. This shows that, when the AI correctly identifies the location of the stricture, its therapeutic suggestion tends to be adequate and compatible with standard clinical conduct.

Additionally, it is important to acknowledge that the artificial intelligence model evaluated in this study was ChatGPT (version 4.0 – 1.2025.105), selected because it is currently the most widespread and accessible language AI, widely used by health professionals and laypeople worldwide. However, other AI platforms with a dedicated visual architecture or specific training in medical images—such as Gemini, DeepSeek, or integrated multimodal models—may show superior performance in interpreting imaging studies such as VCUG. The choice of ChatGPT, therefore, was based on its popularity, availability, and the study's objective of simulating a realistic clinical use scenario by non-specialists. Future comparisons between different AI models may offer additional insights as to which system presents the best balance between diagnostic accuracy, therapeutic proposal, and clinical applicability.

To date, no published studies were found that specifically assess the performance of ChatGPT in interpreting voiding cystourethrograms (VCUGs) for the diagnosis of urethral stricture. However, recent research demonstrates the potential of artificial intelligence (AI) in similar contexts. For example, one study used deep learning to detect and classify urethral strictures on retrograde urethrograms, achieving an accuracy of 91.53% in identifying and categorizing these strictures [27]. In addition, another study employed a machine learning algorithm on retrograde urethrography images, obtaining a urethral stricture detection rate of 88.5% [28]. These results suggest that AI models specifically trained for medical image analysis may offer superior performance in specific diagnostic tasks. The choice of ChatGPT in this study was motivated by its wide diffusion and accessibility among users, aiming to simulate a realistic clinical use scenario by non-specialist professionals. Nevertheless, it is plausible that other AI platforms, especially those with targeted training for medical images, may present more accurate results in identifying urethral strictures.

It is important to emphasize that AI is constantly evolving [29]. The version used in this study (ChatGPT 4.0 – 1.2025.105) already presents specific limitations for the diagnostic task of VCUGs, but newer versions with expanded visual capabilities, improved contextual memory, and greater multimodal integration (such as models that unify text and image) may show better performance in a short period of time. Therefore, this study represents an initial evaluation milestone, and its future replication with more up-to-date versions is highly recommended.

Finally, the study's limitations include: analysis performed with static images rather than dynamic sequences; absence of clinical data in the AI's analysis; exclusion of other stricture topographies; and the model's limitation in correlating multiple images from the same exam. Despite this, the findings reinforce the feasibility of using AI as a medical decision-support tool, with potential for future expansion as models are improved and trained based on image datasets specific to reconstructive urology.

Conclusion

This study demonstrated that the ChatGPT-4 AI model showed high sensitivity and accuracy in detecting bulbar stricture on VCUGs, but with low specificity, suggesting a limitation in differentiating normal exams. Despite failures in the analysis of cases without stricture, the AI showed a good ability to propose therapeutic plans aligned with clinical practice, reinforcing its potential as an auxiliary tool, especially in locations with a shortage of specialists.

This is the first study, as far as is known, to evaluate the use of ChatGPT for interpreting VCUG in urethral strictures. Although the results are promising, clinical use of AI still demands caution. More specialized or computer-vision–based models may offer superior performance and should be explored in future research.

Acknowledgments

We thank the Urology Department of IAMSPE-SP for institutional support.

References

1. Palminteri E, Berdondini E, Verze P, et al. Contemporary urethral stricture characteristics in the developed world. Urology. 2013; 81: 191-197.
2. Joshi P, Kaya C, Kulkarni S, et la. Approach to bulbar urethral strictures: which technique and when?. Turk J Urol. 2016; 42: 53-59.
3. Alberca-Del Arco F, Santos-Pérez De La Blanca R, Amores Vergara C, et al. Bulbar urethroplasty techniques and stricture recurrence: end-to-end versus graft. Minerva Urol Nephrol. 2024; 76: 563-572.
4. Wessells H, Morey A, Souter L, et al. Urethral stricture disease guideline amendment (2023). J Urol. 2023; 210: 64-71.
5. Goulao B, Carnell S, Shen J, et al. Surgical Treatment for Recurrent Bulbar Urethral Stricture: A Randomised Open-label Superiority Trial of Open Urethroplasty Versus Endoscopic Urethrotomy (the OPEN Trial). Eur Urol. 2020; 78: 572-580.
6. Rammah AM, Ghoneima W, Sabry AE, et al. The outcome of nontransecting anastomotic urethroplasty in recurrent bulbar urethral stricture and its impact on sexual functions: A prospective observational study. Urol Sci. 2024; 35: 196-201.
7. Liao RS, Erica Stern, James E Wright, et al. Contemporary management of bulbar urethral strictures. Rev Urol. 2020; 22: 139-151.
8. Lumen N, Hoebeke P, Willemsen P, et al. Etiology of urethral stricture disease in the 21st century. J Urol. 2009; 182: 983-987.
9. Astolfi RH, Lebani BR, Krebs RK, et al. Specific characteristics of urethral strictures in a developing country (Brazil). World J Urol. 2019; 37: 661-666.
10. Abhulimen V, Ofuru VO. Aetiopathogenesis of urethral stricture disease in a tertiary hospital in Southern Nigeria. Int Surg J. 2022; 10: 11-17.
11. Freitas PS, Alves AS, Correia PS, et al. Urethrocystography: a guide for urological surgery?. Diagn Interv Radiol. 2023; 29: 9-17.
12. Dabela-Biketi A, Mawad K, Li H, et al. Urethrographic Evaluation of Anatomic Findings and Complications after Perineal Masculinization and Phalloplasty in Transgender Patients. RadioGraphics. 2020; 40: 393-402.
13. Angermeier KW, Rourke KF, Dubey D, et al. SIU/ICUD Consultation on urethral strictures: evaluation and follow-up. Urology. 2014; 83: S8-S17.
14. Gelman J, Furr J. Urethral stricture disease: evaluation of the male urethra. J Endourol. 2020; 34: S2-S6.
15. Kuo TLC, Venugopal S, Inman RD, et al. Surgical tips and tricks during urethroplasty for bulbar urethral strictures focusing on accurate localisation of the stricture: results from a tertiary centre. Eur Urol. 2015; 67: 764-770.
16. Sheehan JL, Naringrekar HV, Misiura AK, et al. The pre-operative and post-operative imaging appearances of urethral strictures and surgical techniques. Abdom Radiol (NY). 2021; 46: 2115-2126.
17. Harris D, Zhou C, Girardot J, et al. Imaging in urethral stricture disease: an educational review of current techniques with a focus on MRI. Abdom Radiol (NY). 2023; 48: 1062-1078.
18. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. NPJ Digit Med. 2021; 4: 5.
19. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer?. Am J Med. 2018; 131: 129-133.

20. Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. World J Gastroenterol. 2019; 25: 672-682.

21. Coppola F, Faggioni L, Gabelloni M, et al. Human, All Too Human? An All-Around Appraisal of the "Artificial Intelligence Revolution" in Medical Imaging. Front Psychol. 2021; 12: 710982.

22. Chu KY, Tradewell MB. Artificial intelligence in urology. Artificial Intelligence in Medicine. 2021.

23. Shaker G. 175 Beyond human capabilities: the potential of AI in robotic urology. Br J Surg. 2024; 111: znae163.721.

24. Sherazi A, Canes D. Comprehensive analysis of the performance of GPT-3.5 and GPT-4 on the American Urological Association self-assessment study program exams from 2012-2023. Can Urol Assoc J. 2023.

25. Touma NJ, Patel R, Skinner T, et al. AI as a discriminator of competence in urological training: Are We There?. J Urol. 2025; 213: 504-511.

26. McCallum RW, Colapinto V. The role of urethrography in urethral disease. Part I. Accurate radiological localization of the membranous urethra and distal sphincters in normal male subjects. J Urol. 1979; 122: 607-611.

27. Gurung N, Udaya Kumar L, Chandrasekar SN, et al. Deep learning based detection of urethral stricture: segmentation & classification. medRxiv. 2024.

28. Kim JK, McCammon K, Robey C, et al. Identifying urethral strictures using machine learning: a proof-of-concept evaluation of convolutional neural network model. World J Urol. 2022; 40: 3107-3111.

29. Joshi MA. The advancement of artificial intelligence. SSRN Journal. 2024.