Clinical Investigation / Araştırma

DOI: 10.4274/tjod.galenos.2026.99602

Turk J Obstet Gynecol

Diagnostic performance of ChatGPT in adnexal mass classification: A comparative study with frozen section and final histopathology

Adneksiyal kitlelerin sınıflandırılmasında ChatGPT'nin tanısal performansı: Frozen kesit ve nihai histopatoloji ile karşılaştırmalı bir çalışma

Eda Adeviye Şahin[1], Turan Şahin[2], Sonat Erol[2], Ali Konu[3], Savaş Özgür Ağlamış[4], Arzu Bilge Tekin[2], Hanifi Şahin[5]

[1]Eda Şahin Clinic, İstanbul, Türkiye

[2]University of Health Sciences Türkiye, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital, Department of Obstetrics and Gynecology, İstanbul, Türkiye

[3]University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Department of Obstetrics and Gynecology, İstanbul, Türkiye

[4]Özgür Ağlamış Clinic, İstanbul, Türkiye

[5]İstanbul Health and Technology University, İstanbul, Türkiye

Abstract

Objective: To evaluate the diagnostic performance of ChatGPT in classifying adnexal masses as benign, borderline, or malignant, and to compare its performance with intraoperative frozen-section analysis, using final histopathology as the reference standard.

Materials and Methods: This retrospective, observational, comparative study included 259 patients who underwent surgery for ovarian masses between January 2021 and July 2025. Preoperative clinical data, laboratory parameters, and ultrasonographic findings were recorded. ChatGPT assessments were performed using structured clinical inputs without prompt optimization. Frozen section diagnoses and ChatGPT classifications were compared with final histopathological results. Diagnostic performance metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), Cohen's kappa coefficient, and area under the curve (AUC), were calculated.

Results: According to final histopathology, 83.4% of cases were benign, 7.7% were borderline, and 8.9% were malignant. Frozen section analysis demonstrated high diagnostic accuracy, with sensitivity and specificity of 94.4% and 100%, respectively, when borderline cases were excluded. In comparison, ChatGPT showed a sensitivity of 90.0% and a specificity of 92.5%. The PPV for ChatGPT (56.3%) was markedly lower than that for frozen section (100%), whereas NPV values were comparable (98.9% vs. 99.5%). When borderline tumors were classified as malignant, ChatGPT's performance declined further, with a sensitivity of 74.4%, a specificity of 79.6%, and a PPV of 42.1%. Receiver operating characteristic analysis showed AUCs of 0.967 for the frozen section and 0.893 for ChatGPT. ChatGPT demonstrated a higher rate of false-positive classifications, particularly in the borderline and malignant categories.

Conclusion: Frozen section analysis remains a highly reliable method for intraoperative evaluation of ovarian tumors. Although ChatGPT demonstrates good discriminative ability and high NPV, its lower specificity and PPV limit its role as a standalone diagnostic tool. ChatGPT may serve as a complementary decision-support tool, particularly in ruling out malignancy, but cannot replace histopathological evaluation.

Keywords: Ovarian neoplasms, adnexal masses, frozen section, artificial intelligence, ChatGPT, diagnostic accuracy

PRECIS: ChatGPT demonstrated promising diagnostic performance in adnexal mass classification but remained inferior to frozen section, highlighting its potential as a complementary rather than standalone diagnostic tool.

Corresponding Author/Sorumlu Yazar: Turan Şahin MD,
University of Health Sciences Türkiye, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital, Department of Obstetrics and Gynecology, İstanbul, Türkiye
E-mail: dr.turanshn@gmail.com **ORCID ID:** orcid.org/0000-0003-1616-9735

Received/Geliş Tarihi: 24.08.2026 **Accepted/Kabul Tarihi:** 09.11.2026 **Epub:** 12.11.2026

Cite this article as: Şahin EA, Şahin T, Erol S, Konu A, Ağlamış SÖ, Tekin AB, et al. Diagnostic performance of ChatGPT in adnexal mass classification: a comparative study with frozen section and final histopathology. Turk J Obstet Gynecol. [Epub Ahead of Print]

Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of Turkish Society of Obstetrics and Gynecology.
This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Öz

Amaç: Adneksiyal kitlelerin benign, borderline veya malign olarak sınıflandırılmasında ChatGPT'nin tanısal performansını değerlendirmek ve nihai histopatoloji referans standart kabul edilerek performansını intraoperatif frozen kesit analizi ile karşılaştırmaktır.

Gereç ve Yöntemler: Bu retrospektif, gözlemsel ve karşılaştırmalı çalışmaya Ocak 2021 ile Temmuz 2025 tarihleri arasında over kitlesi nedeniyle cerrahi uygulanan 259 hasta dahil edildi. Preoperatif klinik veriler, laboratuvar parametreleri ve ultrasonografik bulgular kaydedildi. ChatGPT değerlendirmeleri, *prompt* optimizasyonu yapılmaksızın yapılandırılmış klinik girdiler kullanılarak gerçekleştirildi. Frozen kesit tanıları ve ChatGPT sınıflandırmaları nihai histopatolojik sonuçlarla karşılaştırıldı. Duyarlılık, özgüllük, pozitif prediktif değer (PPV), negatif prediktif değer (NPV), Cohen kappa katsayısı ve eğri altında kalan alan (AUC) dahil olmak üzere tanısal performans ölçütleri hesaplandı.

Bulgular: Nihai histopatoloji sonuçlarına göre olguların %83,4'ü benign, %7,7'si borderline ve %8,9'u malign olarak saptandı. Borderline olgular dışlandığında frozen kesit analizi sırasıyla %94,4 duyarlılık ve %100 özgüllük ile yüksek tanısal doğruluk gösterdi. Buna karşılık ChatGPT'nin duyarlılığı %90,0, özgüllüğü ise %92,5 olarak bulundu. PPV ChatGPT'de (%56,3), frozen kesite (%100) kıyasla belirgin olarak daha düşükken, NPV'ler benzerdi (sırasıyla %98,9 ve %99,5). Borderline tümörler malign olarak sınıflandırıldığında ChatGPT'nin performansı daha da azalarak duyarlılık %74,4, özgüllük %79,6 ve PPV %42,1 olarak saptandı. Alıcı işletim karakteristiği analizinde AUC değeri frozen kesit için 0,967, ChatGPT için ise 0,893 olarak bulundu. ChatGPT, özellikle borderline ve malign kategorilerde daha yüksek oranda yanlış pozitif sınıflandırma gösterdi.

Sonuç: Frozen kesit analizi, over tümörlerinin intraoperatif değerlendirilmesinde oldukça güvenilir bir yöntem olmaya devam etmektedir. ChatGPT iyi bir ayırt edici performans ve yüksek NPV göstermesine rağmen, daha düşük özgüllüğü ve PPV bağımsız bir tanı aracı olarak kullanımını sınırlamaktadır. ChatGPT, özellikle malignitenin dışlanmasına yardımcı olmak amacıyla tamamlayıcı bir klinik karar destek aracı olarak kullanılabilir; ancak histopatolojik değerlendirmenin yerini alamaz.

Anahtar Kelimeler: Over neoplazmları, adneksiyal kitleler, frozen kesit, yapay zeka, ChatGPT, tanısal doğruluk

Introduction

Ovarian tumors represent a heterogeneous group of gynecological pathologies, encompassing a broad spectrum from benign to borderline and malignant lesions. Among these, ovarian cancer is associated with the highest mortality rate among gynecologic cancers, largely due to late-stage diagnosis and challenges in early detection[1,2]. Therefore, accurate preoperative and intraoperative assessment is crucial for optimal surgical management and improved patient outcomes.

Preoperative evaluation of adnexal masses typically relies on a combination of ultrasonographic findings, serum tumor markers, and clinical characteristics. Structured models, particularly those developed by the International Ovarian Tumor Analysis (IOTA) group, have significantly improved the classification of adnexal masses and risk stratification[3]. However, a definitive diagnosis is usually established through histopathological examination.

Intraoperative frozen section analysis plays a pivotal role in guiding surgical decision-making and has demonstrated high diagnostic accuracy in differentiating benign from malignant ovarian tumors. Nevertheless, its performance may be limited by factors such as sampling error, pathologist experience, and inherent diagnostic challenges, especially in borderline tumors[4,5].

In recent years, advances in artificial intelligence (AI) have led to increasing interest in the use of large language models (LLMs) in clinical decision support. In this context, ChatGPT has emerged as a novel tool capable of processing and interpreting complex clinical data. Although preliminary studies suggest that such models may achieve high accuracy in certain medical tasks, their reliability in complex, multi-parameter clinical decision-making remains uncertain[6]. The evaluation of ovarian tumors requires the integration of multiple clinical, laboratory, and imaging parameters. Whether AI-based systems can effectively synthesize these data and assist in clinical decision-making remains an important research question.

The aim of this study was to evaluate the diagnostic performance of ChatGPT in classifying adnexal masses as benign, borderline, or malignant, and to compare its performance with intraoperative frozen section analysis using final histopathology as the reference standard.

Materials and Methods

This retrospective, observational, comparative study was conducted at the Department of Obstetrics and Gynecology, University of Health Sciences Türkiye, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital. Ethical approval was obtained from the University of Health Sciences Türkiye, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital Scientific Research Ethics Committee (approval number: 2025/510, date: 24.12.2025). The study was carried out in accordance with the principles of the Declaration of Helsinki. Due to the retrospective nature of the study, informed consent was waived, and all data were anonymized prior to analysis.

Patients who underwent surgery for ovarian masses between January 2021 and July 2025 and had intraoperative frozen-section evaluation were assessed for eligibility. Intraoperative frozen section analysis was performed on patients with adnexal masses when malignancy could not be reliably excluded on preoperative clinical, laboratory, and ultrasonographic assessment, or when suspicious findings were identified intraoperatively. Patients who had received prior chemotherapy for ovarian malignancy and those with known unresectable metastatic disease were not included in the study cohort. Only cases with available final histopathological diagnoses were considered eligible. A total of 311 patients were initially assessed for eligibility. Of these, 21 patients were excluded because of incomplete patient records, 28 were excluded because of incomplete clinical data, and 3

were excluded because of insufficient tissue samples or technical artifacts. Thus, a total of 52 patients were excluded, and the remaining 259 patients were included in the final analysis.

Data were retrospectively obtained from the hospital's electronic medical record system. For each patient, demographic characteristics, preoperative laboratory parameters, and ultrasonographic findings were recorded. Laboratory parameters included cancer antigen 125 (CA-125), cancer antigen 19-9 (CA 19-9), carcinoembryonic antigen (CEA), alpha-fetoprotein (AFP), and cancer antigen 15-3 (CA 15-3). Ultrasonographic evaluation included tumor laterality (right, left, or bilateral), tumor size, echogenicity, internal structure (cystic, solid, or mixed), border characteristics (regular or irregular), and presence or absence of ascites. In addition, intraoperative frozen section diagnoses and final histopathological results were documented. All data were anonymized and coded, and no patient-identifiable information was included in the dataset.

A single investigator performed ChatGPT assessments between August and December 2025 using the publicly available free version of ChatGPT (GPT-5.1; OpenAI, San Francisco, CA, USA). For each case, a structured set of clinical inputs, including laboratory parameters and ultrasonographic findings, was entered into the model. All cases were evaluated using a consistent data format, and the model was asked to classify each adnexal mass as benign, borderline, or malignant. The standardized prompt template used for all ChatGPT assessments is provided in Appendix 1. No prompt optimization, iterative querying, or output refinement were performed. This approach was designed to reflect real-world clinical usage rather than an optimized AI pipeline.

The final histopathological diagnosis was accepted as the reference standard for all analyses. The diagnostic performance of frozen-section and ChatGPT classifications was evaluated against this reference. The ability of each method to distinguish between benign, borderline, and malignant lesions was assessed separately.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, version 28.0 (IBM Corp., Armonk, NY, USA). Continuous variables were expressed as mean ± standard deviation or median (interquartile range), depending on data distribution. Categorical variables were presented as frequencies and percentages. Diagnostic performance metrics, including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), were calculated. Agreement between diagnostic methods was assessed using Cohen's kappa coefficient. Receiver operating characteristic (ROC) curve analysis was conducted to evaluate the discriminative ability of each method, and the area under the curve (AUC) was calculated. A p-value of <0.05 was considered statistically significant.

Results

A total of 259 patients were included in the study. The baseline demographic, clinical, and tumor characteristics of the study population are summarized in Table 1. The mean age of the patients was 47.0±14.2 years, and the mean body mass index was 27.5±4.6 kg/m^2. The median gravidity and parity were 3 (1-4) and 2 (1-3), respectively. Among comorbidities, hypertension (27.4%) and diabetes mellitus (15.1%) were the most common. Abdominal pain was present in 66.0% of patients, while menstrual irregularity was observed in 12.4%.

Table 1. Baseline demographic, clinical, and tumor characteristics of the study population

Variable	Mean ± SD/ median (IQR)	Min-max
Age (years)*	47.0±14.2	19-85
Body mass index (kg/m^2)*	27.5±4.6	17.6-39.6
Parity status**		
Gravidity	3.0 (1.0-4.0)	0-14
Parity	2.0 (1.0-3.0)	0-14
Comorbidities, n (%)		
Hypertension	71 (27.4%)	
Diabetes mellitus	39 (15.1%)	
Coronary artery disease	4 (1.5%)	
Chronic kidney disease	3 (1.2%)	
Thyroid disorders	25 (9.7%)	
Abdominal pain, n (%)	171 (66.0%)	
Menstrual irregularity, n (%)	32 (12.4%)	
Side, n (%)		
Right	126 (48.6%)	
Left	112 (43.2%)	
Bilateral	21 (8.1%)	
Tumor size (cm)**		
Right	6.5 (4.0-11.5)	3-32
Left	7.2 (5.9-10.3)	3-25
Tumor structure, n (%)		
Cystic	173 (66.8%)	
Solid	14 (5.4%)	
Cystic/solid	72 (27.8%)	
Echogenicity, n (%)		
Anechoic	65 (25.1%)	
Hyperechoic	48 (18.5%)	
Hypoechogenic	62 (23.9%)	
Hyper/hypoechoic	84 (32.4%)	
Presence of ascites, n (%)	19 (7.3%)	
Border irregularity, n (%)	18 (6.9%)	

*Mean ± SD, **median (IQR)
SD: Standard deviation, IQR: Interquartile range

Regarding tumor characteristics, 48.6% of lesions were located in the right ovary, 43.2% in the left ovary, and 8.1% were bilateral. The median tumor size was 6.5 cm (4.0-11.5) for right-sided lesions and 7.2 cm (5.9-10.3) for left-sided lesions. In terms of structural composition, the majority of tumors were cystic (66.8%), followed by mixed cystic-solid lesions (27.8%) and solid lesions (5.4%). Echogenicity was heterogeneous, with the most common pattern being mixed hyper/hypoechoic (32.4%). Ascites was present in 7.3% of patients, and irregular tumor borders were identified in 6.9% of patients (Table 1).

Laboratory parameters are presented in Table 2. The mean preoperative hemoglobin level was 12.3±1.4 g/dL, while the mean postoperative hemoglobin level was 10.8±1.3 g/dL. The median values of tumor markers were as follows: CA-125, 21.7 U/mL (12.8-40.5); CA 19-9, 10.8 U/mL (7.0-38.0); CEA, 1.6 ng/mL (1.2-2.0); AFP, 2.2 ng/mL (1.5-4.8); and CA 15-3, 16.5 U/mL (8.2-22.5). A wide range of values (minimum to maximum) was observed across all parameters, particularly for CA-125 and CA 19-9, indicating considerable variability (Table 2).

According to final histopathology, 216 cases (83.4%) were benign, 20 cases (7.7%) were borderline, and 23 cases (8.9%) were malignant. Among 217 cases classified as benign in the frozen-section analysis, 211 (97.2%) were confirmed as benign, 5 (2.3%) were borderline, and 1 (0.5%) was malignant. Among the 25 cases classified as borderline, 60.0% remained borderline, whereas 20.0% were benign and 20.0% were malignant. All 17 cases classified as malignant by frozen section were confirmed on final pathology (Table 3).

Among 183 cases classified as benign in the ChatGPT-based evaluation, 172 (94.0%) were truly benign, 9 (4.9%) were borderline, and 2 (1.1%) were malignant. Among the 38 cases classified as borderline, 78.9% were benign, only 13.2% were borderline, and 7.9% were malignant. Among the 38 cases classified as malignant, only 47.4% were confirmed malignant, while 36.8% were benign and 15.8% were borderline. These findings indicate that ChatGPT produced a higher rate of false-positive classifications, particularly in the borderline and malignant categories (Table 3).

Table 2. Descriptive statistics of laboratory parameters

Variable	Mean ± SD/ median (IQR)	Min-max
Hemoglobin (preoperative)*	12.3±1.4	8.3-16.2
Hemoglobin (postoperative)*	10.8±1.3	7.2-14.4
CA-125 (U/mL)**	21.7 (12.8-40.5)	2.4-4222.0
CA 19-9 (U/mL)**	10.8 (7.0-38.0)	0.8-1569.0
CEA (ng/mL)**	1.6 (1.2-2.0)	0.3-199.0
AFP (ng/mL)**	2.2 (1.5-4.8)	0.3-1060.0
CA 15-3 (U/mL)**	16.5 (8.2-22.5)	1.0-385.0

*Mean ± SD, **median (IQR)
SD: Standard deviation, IQR: Interquartile range, CEA: Carcinoembryonic antigen, AFP: Alpha-fetoprotein, CA-125: Cancer antigen 125, CA 19-9: Cancer antigen 19-9, CA 15-3: Cancer antigen 15-3

In the diagnostic performance analysis excluding borderline cases, the sensitivity and specificity of the frozen section were 94.4% and 100%, respectively. For ChatGPT, sensitivity was 90.0% and specificity was 92.5%. The PPV was 100% for frozen section and 56.3% for ChatGPT, while the NPV was 99.5% for frozen section and 98.9% for ChatGPT. Cohen's kappa coefficient indicated near-perfect agreement for the frozen section (κ=0.969) and moderate agreement for ChatGPT (κ=0.651) (Table 4a).

When borderline tumors were classified as malignant, the sensitivity and specificity of the frozen section were 86.0% and 97.7%, respectively. For ChatGPT, these values decreased to 74.4% and 79.6%. The PPV was 88.1% for frozen section and 42.1% for ChatGPT, and the NPV was 97.2% for frozen section and 94.0% for ChatGPT. Cohen's kappa coefficient was 0.826 for the frozen section and 0.393 for ChatGPT, indicating a marked decline in agreement under this classification approach (Table 4b).

Table 3. Distribution of diagnostic classifications according to frozen section, ChatGPT, and final pathology

	Final pathology			
	Benign	Borderline	Malign	Total
Frozen				
Benign	211 (97.2%)	5 (2.3%)	1 (0.5%)	217
Borderline	5 (20.0%)	15 (60.0%)	5 (20.0%)	25
Malign	0	0	17 (100%)	17
ChatGPT				
Benign	172 (94.0%)	9 (4.9%)	2 (1.1%)	183
Borderline	30 (78.9%)	5 (13.2%)	3 (7.9%)	38
Malign	14 (36.8%)	6 (15.8%)	18 (47.4%)	38
Total	216	20	23	

Table 4a. Diagnostic performance of frozen section and ChatGPT (borderline cases excluded)

	Frozen	ChatGPT
Sensitivity	94.4%	90.0%
Specificity	100%	92.5%
PPV	100%	56.3%
NPV	99.5%	98.9%
Kappa	0.969	0.651

Table 4b. Diagnostic performance when borderline tumors are classified as malignant

	Frozen	ChatGPT
Sensitivity	86.0%	74.4%
Specificity	97.7%	79.6%
PPV	88.1%	42.1%
NPV	97.2%	94.0%
Kappa	0.826	0.393

PPV: Positive predictive value, NPV: Negative predictive value

ROC curve analysis demonstrated excellent discriminative ability for frozen section, with an AUC of 0.967. ChatGPT also showed good discriminative performance, with an AUC of 0.893; however, its performance remained inferior to that of the frozen section (Figure 1).

Discussion

Accurate preoperative and intraoperative classification of ovarian tumors is critical for determining the extent of surgery, planning fertility-sparing approaches, and avoiding unnecessarily extensive surgical procedures. Epithelial ovarian cancer continues to represent a significant oncological burden in women, with notably high mortality, particularly in cases diagnosed at advanced stages[7]. Contemporary reviews also emphasize that the evaluation of ovarian masses requires the integrated interpretation of clinical, biochemical, and imaging data, and that decisions based on a single parameter are often insufficient[8].

The main finding of this study was that frozen section analysis demonstrated very high diagnostic accuracy compared with final histopathology, whereas ChatGPT exhibited lower specificity and PPV, particularly because it tended to misclassify benign cases as borderline or malignant. When borderline cases were excluded, the sensitivity, specificity, kappa, and AUC for the frozen section were 94.4%, 100%, 0.969, and 0.967, respectively; for ChatGPT, they were 90.0%, 92.5%, 0.651, and 0.893. When borderline tumors were classified as malignant, the decline in ChatGPT performance became more pronounced, with the kappa coefficient decreasing to 0.393 and PPV to 42.1%. These findings indicate that although ChatGPT performs relatively well in ruling out malignancy, it has clear limitations in confirming malignancy.

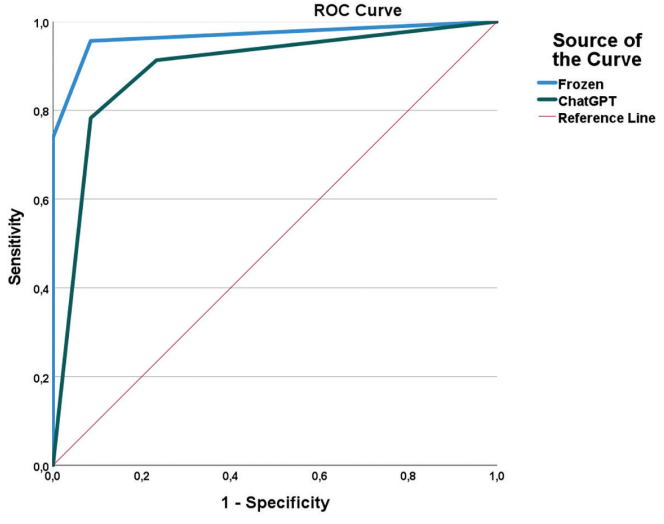


Figure 1. Receiver operating characteristic (ROC) curves of frozen section and ChatGPT for the diagnosis of malignant ovarian tumors

The high performance of the frozen section observed in our study is strongly consistent with the existing literature. Standardized risk classification systems for ovarian adnexal lesions, particularly O-RADS US v2022, aim to provide more consistent reporting of morphological findings and more systematic assessment of malignancy risk[9]. In addition, recent multicenter data have demonstrated that O-RADS-based classification can significantly influence surgical decision-making and reduce unnecessary resections[10]. In our study, the very high specificity and AUC for the frozen section support the notion that intraoperative sampling and morphological interpretation by the pathologist remain among the most reliable components of clinical decision-making.

Nevertheless, the success of modern imaging-based risk systems and IOTA-based models demonstrates that accurate diagnosis depends not only on individual "alarm features," but also on the contextual weighting of these findings. It has been reported that the ADNEX and IOTA two-step strategy performs well in differentiating benign and malignant lesions in low-risk populations[11]. Similarly, validation studies in Japanese women have shown that the ADNEX model is effective in differential diagnosis, although interpretation of features such as solid components and papillary projections may be influenced by the level of experience and training[12]. From this perspective, it is not surprising that ChatGPT did not reach the performance level of frozen section despite receiving structured clinical inputs, as the model may be able to "read" parameters but not to weigh them at an expert level.

Despite the strong performance of the frozen section, borderline tumors have long presented significant diagnostic challenges. A recent pathology series demonstrated that frozen section diagnosis provides high accuracy and reproducibility overall, but the greatest difficulty arises in the borderline category[13]. A contemporary study from a tertiary center in London also reported that borderline ovarian tumors constitute the most problematic group across ultrasound, magnetic resonance imaging, and frozen section methods[14]. Another recent single-center study showed that while frozen section is generally highly useful, sensitivity decreases significantly in borderline tumors[15]. In our study, the deterioration in ChatGPT's performance when borderline cases were included reflects the biological and morphological uncertainty inherent in this category; in other words, the issue lies not only with the model but also with the diagnostic category itself.

One of the most notable weaknesses of ChatGPT in our study was an increase in false-positive classifications. As shown in Table 3, a substantial proportion of cases classified as borderline or malignant by ChatGPT was found to be benign on final pathology. This may be explained by the model assigning excessive weight to features associated with high malignancy risk, such as ascites, irregular borders, mixed structure, and heterogeneous echogenicity. However, these findings alone do not indicate malignancy; accurate interpretation requires

evaluation of these features collectively within the clinical context. In our dataset, ascites was present in only 7.3% of cases and irregular borders were present in 6.9% of cases, yet the model's tendency toward overestimation of malignancy supports this hypothesis.

This finding is also consistent with the general behavioral pattern of LLMs in healthcare. It has been shown that LLMs are capable of encoding clinical knowledge and can achieve impressive performance in many tasks[16]. However, recent evaluations summarizing the working principles of ChatGPT emphasize that while these systems are highly effective in capturing language patterns, they do not always process clinical context with the same depth[17]. Another evaluation in the context of neurological care has reported that such models exhibit variability, limited clinical reasoning capacity, and reproducibility issues[18]. Therefore, the "sensitive but overly cautious" profile observed for ChatGPT in our study is consistent with the broader literature on the safe clinical use of LLMs.

Recent high-impact studies have demonstrated that the performance of LLMs in clinical decision-making depends not only on the knowledge level but also on prompting strategy, sampling approach, and evaluation framework. A comprehensive study published in Nature Medicine showed that LLMs have notable limitations in clinical decision-making but that their performance can be improved with structured evaluation and error-mitigation strategies[19]. Another study published in the Journal of the American Medical Informatics Association reported that reasoning strategies can influence outcomes in medical question-answering tasks, although these improvements are not uniform across all case types[20]. The use of a free version, a single user, and a non-optimized, real-world prompting approach in our study allowed performance to be deliberately measured under "natural clinical use" conditions, thereby providing a more realistic framework than that of idealized AI performance in controlled settings.

One of the notable strengths of our study is that ChatGPT was not provided with vague free-text input; instead, structured ultrasonographic parameters—such as laterality, size, echogenicity, internal structure, border characteristics, and presence of ascites—were systematically entered alongside tumor markers. This is important because recent studies have shown that the inclusion of laboratory data and structured clinical details can improve the diagnostic accuracy of LLM-based differential diagnosis[21]. In addition, it has been reported that GPT-4-level models can assist in complex diagnostic scenarios, although they still do not replace expert clinical reasoning[22]. Therefore, our findings are particularly valuable for demonstrating that, even with structured clinical input, ChatGPT remains inferior to the frozen section.

It is now well established that the performance of LLMs varies according to model version. A retrospective analysis conducted in an emergency department setting reported that GPT-4 achieved diagnostic accuracy comparable to, and in some cases exceeding, that of physicians[23]. However, more recent studies comparing different model generations have demonstrated significant differences in clinical reasoning and guideline adherence across versions[24]. In this context, the findings obtained in our study using a freely available version are historically relevant, as they reflect the performance of LLM technology under accessible, non-standardized, real-world conditions. Re-evaluation of the same dataset using more advanced and fixed model versions would be highly valuable in demonstrating the magnitude of the performance improvement.

The most important clinical message of this study is that ChatGPT represents a promising but currently complementary tool in the assessment of malignancy risk in ovarian masses. The high NPV suggests that the model may provide reassurance in certain cases suggesting benignity; however, the low PPV and marked decline in performance in borderline cases indicate that it is not suitable as a standalone decision-making tool. In other words, ChatGPT may function as a "second opinion," an educational aid, or a structured preliminary risk assessment tool within the diagnostic workflow, but it cannot replace frozen section or final histopathological evaluation. This interpretation is consistent both with our findings and the broader literature, which indicates that LLMs serve as supportive yet limited tools in clinical decision-making.

Study Limitations

This study has several limitations. First, its retrospective, single-center design may limit the generalizability of the findings to different patient populations and clinical settings. Second, benign lesions constituted the majority of the study population, whereas borderline and malignant tumors were relatively few. Only 20 borderline tumors were identified in the final histopathological cohort. This relatively small number may have resulted in greater statistical uncertainty and reduced precision of the diagnostic performance estimates for borderline tumors. Therefore, the findings related to this subgroup should be interpreted with caution and require validation in larger cohorts with a greater representation of borderline tumors. Third, a single investigator performed ChatGPT assessments using the freely available version of the model and a standardized input approach. Since the performance of LLMs may vary depending on the model version, updates, and prompting strategy, our findings may not be directly generalizable to other versions of ChatGPT or different prompting methods. In addition, ChatGPT did not directly evaluate the ultrasonographic images but made its classifications based on predefined ultrasonographic features, laboratory results, and clinical data. Therefore, potentially relevant diagnostic information contained within the original images may not have been fully captured. Finally, ChatGPT was not directly compared with established and validated risk assessment systems such as O-RADS or IOTA-ADNEX.

Conclusion

Future studies should use prospective, multicenter designs with more balanced patient populations, fixed model versions, and standardized prompting protocols. Furthermore, direct comparison of ChatGPT outputs with O-RADS, IOTA-ADNEX, and frozen section within the same cohort, and testing hybrid models that integrate these variables, may yield more clinically applicable results. Subgroup analyses focusing on diagnostically challenging scenarios—such as borderline tumors, benign cases with ascites, and complex mixed lesions—will be particularly important in clarifying the true contribution of AI to clinical practice.

Ethics

Ethics Committee Approval: Ethical approval was obtained from the University of Health Sciences Türkiye, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital Scientific Research Ethics Committee (approval number: 2025/510, date: 24.12.2025).

Informed Consent: Due to the retrospective nature of the study, informed consent was waived, and all data were anonymized prior to analysis.

Footnotes

Authorship Contributions

Concept: E.A.Ş., A.K., S.Ö.A., H.Ş., Design: T.Ş., A.K., A.B.T., H.Ş., Data Collection or Processing: T.Ş., S.E., S.Ö.A., H.Ş., Analysis or Interpretation: T.Ş., E.A.Ş., A.B.T., H.Ş., Literature Search: T.Ş., S.E., A.B.T., H.Ş., Writing: T.Ş., E.A.Ş., S.Ö.A., H.Ş.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: Artificial intelligence (AI) tools were used for language editing. All generated content was critically reviewed, revised, and approved by the authors, who accept full responsibility for the final version of the manuscript.

Appendix 1: https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/5ce8e877-267c-48a2-9368-f50644c0170f.pdf

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024;74:12-49.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209-49.
3. Timmerman D, Testa AC, Bourne T, Ameye L, Jurkovic D, Van Holsbeke C, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol. 2008;31:681-90.
4. Medeiros LR, Rosa DD, Edelweiss MI, Stein AT, Bozzetti MC, Zelmanowicz A, et al. Accuracy of frozen-section analysis in the diagnosis of ovarian tumors: a systematic quantitative review. Int J Gynecol Cancer. 2005;15:192-202.
5. Berek JS, Renz M, Kehoe S, Kumar L, Friedlander M. Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. Int J Gynaecol Obstet. 2021;155(Suppl 1):61-85.
6. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11:887.
7. Webb PM, Jordan SJ. Global epidemiology of epithelial ovarian cancer. Nat Rev Clin Oncol. 2024;21:389-400.
8. Caruso G, Weroha SJ, Cliby W. Ovarian cancer: a review. JAMA. 2025;334:1278-91.
9. Strachowski LM, Jha P, Phillips CH, Blanchette Porter MM, Froyman W, Glanc P, et al. O-RADS US v2022: an update from the American College of Radiology's ovarian-adnexal reporting and data system US Committee. Radiology. 2023;308:e230685.
10. Shen L, Sadowski EA, Gupta A, Maturen KE, Patel-Lippmann KK, Zafar HM, et al. The ovarian-adnexal reporting and data system (O-RADS) US score effect on surgical resection rate. Radiology. 2024;313:e240044.
11. Pascual MA, Vancraeynest L, Timmerman S, Ceusters J, Ledger A, Graupera B, et al. Validation of ADNEX and IOTA two-step strategy and estimation of risk of complications during follow-up of adnexal masses in low-risk population. Ultrasound Obstet Gynecol. 2024;64:395-404.
12. Kadooka M, Suemitsu T, Ashimoto K, Takesawa A, Matsui H, Otsuka I, et al. Validation of the IOTA ADNEX model among Japanese women performed by gynecology trainees and ultrasound specialists: a retrospective diagnostic accuracy study. J Ultrasound Med. 2024;43:1857-68.
13. Zaiem F, Deirawan H, Kherallah R, Fehmi O, Jang H, Kim S, et al. Accuracy and reproducibility of frozen section diagnosis in ovarian tumors. Arch Pathol Lab Med. 2022;146:626-31.
14. Mitchell S, Gleeson J, Tiwari M, Bailey F, Gaughran J, Mehra G, et al. Accuracy of ultrasound, magnetic resonance imaging and intraoperative frozen section in the diagnosis of ovarian tumours: data from a London tertiary centre. BJC Rep. 2024;2:50.
15. Goel P, Rathore SB, Sethi N, Gupta S, Kalani A. Diagnostic accuracy of frozen section in ovarian masses: a single centre study. Med J Armed Forces India. 2025;81:151-6.
16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-80. Erratum in: Nature. 2023;620:E19.
17. Briganti G. How ChatGPT works: a mini review. Eur Arch Otorhinolaryngol. 2024;281:1565-9.
18. Moura L, Jones DT, Sheikh IS, Murphy S, Kalfin M, Kummer BR, et al. Implications of large language models for quality and efficiency of neurologic care: emerging issues in neurology. Neurology. 2024;102:e209497.
19. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30:2613-22.
20. Lucas MM, Yang J, Pomeroy JK, Yang CC. Reasoning with large language models for medical question answering. J Am Med Inform Assoc. 2024;31:1964-75.

21. Bhasuran B, Jin Q, Xie Y, Yang C, Hanna K, Costa J, et al. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. NPJ Digit Med. 2025;8:166.

22. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med (Lausanne). 2024;11:1380148.

23. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. J Med Internet Res. 2024;26:e56110.

24. Babaee Hemmati Y, Rasouli M, Falahchai M. Comparative analysis of ChatGPT-3.5 and GPT-4 in open-ended clinical reasoning across dental specialties. Eur J Dent Educ. 2026;30:526-32.