

Evaluating AI Models for Ovarian Mass Classification and Diagnosis Using IOTA Parameters

Omri Dominsky^{1,2}, Aya Mudrik¹, Hadas Miremberg^{1,2}, Ran Nagar^{1,2}, Karina Krajden Haratz^{1,2}, Sharon Perlman^{1,2}

1.Division of Ultrasound in ObGyn at Lis Maternity and Women’s Hospital, Tel Aviv Medical Center, Israel

2. Gray Faculty of Medical & Health Sciences, Tel Aviv University;

Objectives

To evaluate the performance of ChatGPT-4o (cGPT), Claude, and Perplexity in classifying ovarian masses using IOTA criteria compared to expert gynecologists.

Methods

35 TVS images (five per mass type), classified by an expert gynecologist (serving as the reference standard), were analyzed by GPT-4o, Claude, and Perplexity for classification accuracy (benign vs. malignant), correct diagnosis, and IOTA adherence

Results

cGPT achieved the highest performance (62.9% classification, 57.1% diagnosis, 63.2% IOTA adherence), followed by Claude (51.4%, 40%, 56.1%) and Perplexity (40%, 34.3%, 48.2%), with significant variation across cyst types. Simple cysts were best classified as benign (100% accuracy by cGPT and Claude, 60% by Perplexity). Endometrioma showed inconsistencies, as Claude classified 100% correctly but failed in diagnosis (0%, misclassified as simple cyst), while cGPT and Perplexity classified 80% but also failed in diagnosis. Among complex cysts, mucinous neoplasms were best identified (cGPT: 80% classification, 100% diagnosis, 70% IOTA adherence). Ovarian epithelial cancer had moderate AI performance (cGPT, Claude: 60%, Perplexity: 20%). Dermoid cysts and fibromas were frequently misclassified, while borderline tumors had the lowest accuracy (Claude: 0%, cGPT: 20% diagnosis).

Comparison of the performance of ChatGPT-4o, Claude, and Perplexity in classifying ovarian masses based on IOTA criteria.

Mass type	ChatGPT -40			Claude AI			Perplexity		
	Classification	Diagnosis	IOTA Adherence	Classification	Diagnosis	IOTA Adherence	Classification	Diagnosis	IOTA Adherence
Overall	62.9%	57.1%	63.2%	51.4%	40%	56.1%	40%	34.3%	48.2%
Simple Cyst	100%	100%	87.5%	100%	100%	85%	60%	60%	85%
Dermoid Cyst	40%	40%	60%	20%	0	47.5%	0	0	42.5%
Endometrioma	80%	0	52.5%	100%	0	65%	80%	0	55%
Fibroma	40%	0	47.5%	40%	0	42.5%	60%	0	45%
Borderline Tumor	20%	20%	42.5%	0	20%	42.5%	40%	40%	37.5%
Mucinous Neoplasm	80%	100%	70%	40%	60%	40%	80%	80%	50%
Ovarian Cancer	60%	60%	45%	60%	60%	35%	20%	20%	30%

Conclusions: Chat-GPT4o outperformed Claude and Perplexity in analyzing ovarian cyst images but remained less accurate than expert interpretation. While AI demonstrates promise as a diagnostic tool, further refinement is required before clinical implementation, especially for malignant and borderline lesions.