Interdisciplinary Research in Medical Sciences Specialty
ISSN: 3062-4401
2026, Volume 6, Issue 1, Page No: 240-254

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Deep Learning–Based Ultrasound Screening Pipeline for the Diagnosis of Ovarian Masses: A Retrospective Multicenter Validation Study

Mateo Alvarez[1]*, Sofia Herrera[2], Diego Cruz[1], Andres Castro[3]

[1]Department of Ultrasound Imaging and Deep Learning, Faculty of Medicine, National University of Colombia, Bogota, Colombia.
[2]Department of Ovarian Cancer Diagnostics, Faculty of Medicine, University of Chile, Santiago, Chile.
[3]Department of Retrospective Validation Studies, Faculty of Medicine, University of Antioquia, Medellin, Colombia.

***E-mail** ✉ mateo.alvarez@unal.edu.co

Received: 16 March 2026; **Revised:** 28 May 2026; **Accepted:** 05 June 2026

ABSTRACT

Ovarian cancer exhibits the highest death rate among gynecological cancers and is typically first evaluated through ultrasound. However, the intricate nature of ultrasound visuals for ovarian lesions, combined with the deep pelvic positioning, demands substantial expertise for accurate subjective interpretation. Consequently, identifying ovaries and ovarian lesions, along with diagnosing ovarian cancer, remains difficult. This study sought to create an automated deep learning-based system called the Ovarian Multi-Task Attention Network (OvaMTA). It targets ovary and ovarian mass detection, segmentation, and classification, while also supporting the diagnosis of ovarian masses from ultrasound screenings. From June 2020 to May 2022, the OvaMTA model underwent training, validation, and testing using a training/validation group of 6938 images and an internal test group of 1584 images. These were gathered from 21 hospitals among women receiving ultrasound for ovarian masses. Two additional external test groups were then assembled from two other hospitals. This included 1896 images collected from February 2024 to April 2024 for the image-based external test set, plus 159 videos gathered from April 2024 to May 2024 for the video-based external test set. An artificial intelligence (AI) platform named OvaMTA was built to diagnose ovarian masses via ultrasound screening. It comprises two components: OvaMTA-Seg, a whole-image segmentation model for ovary detection, and OvaMTA-Diagnosis, a diagnostic model that predicts the pathological nature of ovarian masses from image patches extracted by OvaMTA-Seg. System performance was assessed across one internal and two external validation sets, with comparisons to physicians' evaluations in practical clinical testing. Eight doctors participated in reviewing the real-world cases. The AI's value in supporting physician diagnoses was also examined. For segmentation tasks, OvaMTA-Seg attained an average Dice score of 0.887 on the internal test set and 0.819 on the image-based external test set. It also showed strong results in detecting ovarian masses across test images (covering both normal ovaries and lesions), with an internal test area under the curve (AUC) of 0.970 and an external test AUC of 0.877. Regarding classification and diagnostic prediction, OvaMTA-Diagnosis delivered strong outcomes on the image-based internal (AUC: 0.941) and external (AUC: 0.941) test sets. In the video-based external evaluation, OvaMTA processed 159 videos containing ovarian masses, achieving an AUC of 0.911. This performance was similar to that of senior radiologists (accuracy [ACC]: 86.2% vs. 88.1%, p = 0.50; sensitivity [SEN]: 81.8% vs. 88.6%, p = 0.16; specificity [SPE]: 89.2% vs. 87.6%, p = 0.68). AI assistance led to notable gains for junior and intermediate radiologists compared to their unaided performance (ACC: 80.8% vs. 75.3%, p = 0.00015; SEN: 79.5% vs. 74.6%, p = 0.029; SPE: 81.7% vs. 75.8%, p = 0.0032). General practitioners supported by AI reached performance levels matching those of radiologists on average (ACC: 82.7% vs. 81.8%, p = 0.80; SEN: 84.8% vs. 82.6%, p = 0.72; SPE: 81.2% vs. 81.2%, p > 0.99). The ultrasound-based OvaMTA system serves as a straightforward and effective support tool for ovarian cancer screening, delivering diagnostic accuracy on par with senior radiologists. It represents a promising resource for ovarian cancer screening efforts.

Keywords: Deep learning, Ovarian cancer, Ovarian mass, Ultrasound, Artificial intelligence pipeline

How to Cite This Article: Alvarez M, Herrera S, Cruz D, Castro A. Deep Learning–Based Ultrasound Screening Pipeline for the Diagnosis of Ovarian Masses: A Retrospective Multicenter Validation Study. Interdiscip Res Med Sci Spec. 2026;6(1):240-54. https://doi.org/10.51847/cSBeqzI731

Introduction

Ovarian cancer ranks among the three most frequent malignant gynecological tumors [1]. It carries the highest mortality rate in this category and poses a major risk to women's health [2]. Due to its deep pelvic location, subtle onset, and lack of distinct early symptoms, it is frequently subject to misdiagnosis or missed diagnosis [3]. Ultrasound (US) imaging is central to ovarian cancer detection. Yet, challenges arising from complex anatomy, variable biological behavior, and the labor-intensive, operator-dependent aspects of US limit standalone diagnostic accuracy [4]. To enhance this, various clinical models and classification systems have been developed for ovarian masses. These draw on patient data, US image characteristics, and lab results, including the Malignant Risk Index (RMI) [5], Gynecological US Imaging Reporting and Data System (GI-RADS) [6], Simple Rules (SR) [7], Attachment Region Lesion Analysis Model (ADNEX) [8], and Ovarian Attachment Reporting and Data System (O-RADS) [9]. Current US-based approaches for ovarian masses still depend heavily on manual feature identification by specialists, leaving room for better accuracy and efficiency. Their limited practicality and broad applicability hinder adoption, especially in primary care settings. Hence, creating a fast and precise automated diagnostic tool for ovarian cancer is essential.

Advances in deep learning (DL) have demonstrated major advantages for US diagnosis by automatically deriving complex image features without manual selection by radiologists. Multiple investigations have tested DL methods on ovarian masses and reported favorable outcomes [10-24]. Nevertheless, a complete, automated AI pipeline that withstands diverse US imaging variations and aligns with routine clinical standards is still missing. Key gaps include: (1) the deep pelvic position of ovaries and masses complicates detection and segmentation, with most existing studies relying on pre-labeled images based on prior knowledge [19, 25], which involves time-consuming manual preprocessing prone to variability; full automation requires combining segmentation and classification; (2) clinical practice involves real-time dynamic video assessment, yet current models focus only on static images; and (3) limited real-time visualization and explainability.

The goal was to build and validate the Ovarian Multi-Task Attention Network (OvaMTA) deep learning pipeline. It enables real-time guidance during US scanning to detect and label normal ovaries and ovarian masses, while screening for ovarian cancer. This used diverse multicenter image datasets for retrospective testing and external validation, with scalability tested on real-world clinical video data. Comparisons included AI-assisted performance of junior radiologists and non-specialist general practitioners versus unaided junior, intermediate, and senior radiologists. Efforts were also made to provide real-time visual interpretation of the model's internal "black box" decisions. The AI platform excelled at localizing tumors and ovaries in US videos and showed outstanding ability to differentiate benign versus malignant ovarian masses. It enhanced diagnostic accuracy for both general practitioners and radiologists, potentially serving as a valuable tool for ovarian cancer identification and diagnosis.

Materials and Methods

Ethics

This retrospective multicenter investigation utilized US images collected from 23 hospitals. Grayscale US images were retrospectively obtained from Zhejiang Provincial People's Hospital and the National Health Commission Medical Imaging Standard Database of China (covering 21 hospitals) between June 2020 and May 2022 for training, validation, and internal testing. The protocol received approval from the Institutional Review Board of Zhejiang Provincial People's Hospital (QT2023117), with a waiver for written informed consent. External test dataset images came from Sichuan Provincial Maternity and Child Health Care Hospital and the Second Affiliated Hospital of Harbin Medical University (February 2024 to April 2024). Videos for external validation dataset B were acquired from the same centers (April 2024 to May 2024). Approvals were granted by the respective Institutional Review Boards: Sichuan Provincial Maternity and Child Health Care Hospital (20240205-011) and the Second Affiliated Hospital of Harbin Medical University (KY2024-029). All participants in the external cohorts provided written informed consent.

Inclusion and exclusion criteria

Inclusion required patients with at least one persistent ovarian mass identified on US. For the training/validation/internal test datasets (June 2020–May 2022), cases either had surgical confirmation with

histopathology (1636 cases) or at least 6 months of US follow-up showing resolution (241 cases). Analysis in this study focused specifically on follicular cysts and corpus luteum cysts. External test dataset A (February–April 2024) and dataset B (April–May 2024) included only cases with pathological confirmation (498 cases total). Images of the contralateral normal ovary (in unilateral mass cases) were also incorporated as normal ovaries. Pathological diagnoses followed the WHO classification of female genital tumours (2020) and were grouped as benign or malignant (including borderline tumors) [26]. Ovarian masses encompassed both ovarian and fallopian tube-origin tumors.

Exclusion criteria covered non-ovarian tumors (except metastatic ovarian tumors verified histologically), cases with unclear histopathology, and those lacking follow-up US. Images with multiple ovarian masses were also excluded. Data collection followed a standardized form by trained staff and was verified by two researchers (Y.N.W. and W.L.D.). A study flowchart appears in **Figure 1**. The research adhered to STARD-2015 guidelines (equator-network.org).

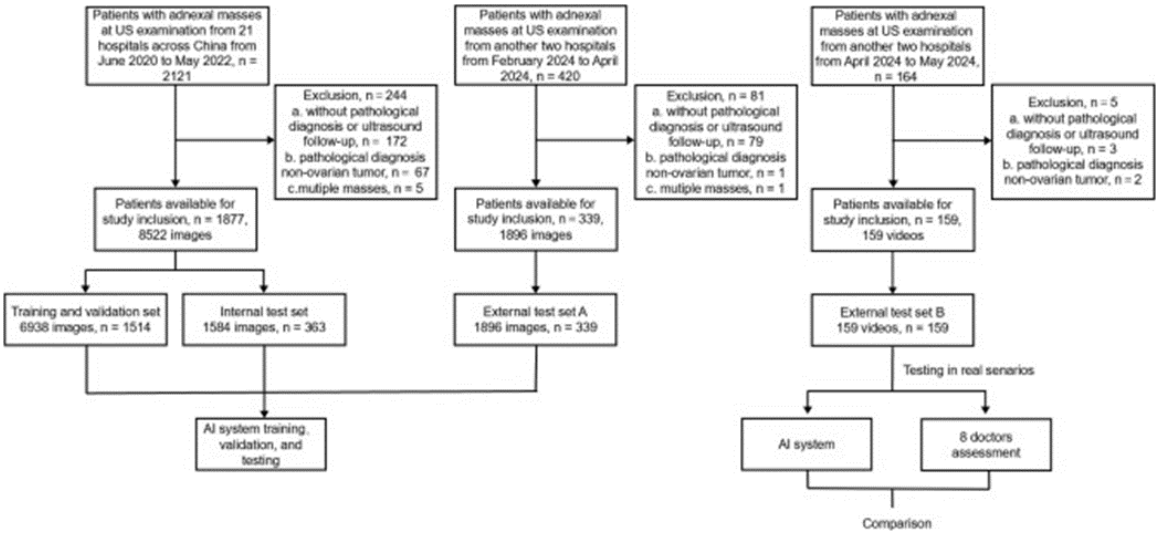


Figure 1. The diagram illustrates the selection standards and the overall process for assessing the AI platform. US = ultrasound; AI = artificial intelligence.

Ultrasound examinations

All ultrasound scans were conducted with 38 different machines from 9 manufacturers, employing transvaginal, transrectal, and transabdominal approaches. Radiologists possessing more than five years of specialized experience in pelvic ultrasonography acquired the preoperative ovarian ultrasound images and recordings according to established protocols. A grayscale dynamic recording of the scan was captured for each ovarian mass at a consistent speed, lasting a minimum of 2–5 seconds. Scans routinely progressed from “no visible structure” to “the largest cross-section” and back to “no visible structure.” The region of interest (ROI) was centered in the frame to produce a uniform ultrasound recording. Stored data for analysis comprised the longitudinal maximum cross-section ultrasound image, the transverse maximum cross-section ultrasound image, and at least one representative image highlighting key features such as papillary projections, posterior acoustic shadowing, or solid components. Ultrasound images of unaffected ovarian tissue from the same patients were also retained for analysis, including at least one longitudinal view and one transverse view.

Training dataset preprocessing

All ovarian ultrasound images were transformed into PNG format. Labeling was performed using a system from Zhejiang Demetics Medical Technology Co., Ltd. (software version 4.4.0.1) to delineate all ovarian masses and normal ovaries. A radiologist with eight years of ultrasound expertise completed the initial outlining and ROI sketching, followed by review and refinement by a senior radiologist with 40 years of experience. JSON files were generated for every dataset, after which OpenCV-Python (version 4.8.1.78) was applied to compute the bounding box around each annotated ROI.

For the OvaMTA-Seg component, which identifies ovaries or nodules across the full image, the complete ultrasound image, binary ROI mask, and pathology labels were utilized directly for supervised training. For

OvaMTA-Diagnosis, to achieve more precise focus on relevant areas, the bounding boxes were expanded by 25 pixels on all sides before cropping the ultrasound images to create patches centered on the ovarian masses. Pathology labels were converted into three-dimensional one-hot encoding, while masks were handled as binary images scaled proportionally to the original image dimensions.

Because images originated from various ultrasound devices, they varied in resolution. All inputs were therefore resized to a uniform 352 × 352 pixels for training and validation. To mitigate overfitting and expand training data variety, the following augmentations were applied to both images and corresponding masks: horizontal flipping (probability 0.5), vertical flipping (probability 0.5), and random clockwise rotation up to 90° [27, 28].

Model architecture and training

The overall pipeline, termed OvaMTA, integrates OvaMTA-Seg and OvaMTA-Diagnosis. OvaMTA-Seg performs initial segmentation of ROIs within ultrasound images and generates a probability score indicating the presence of ovarian masses. In cases without detectable masses, the system outputs the OvaMTA-Seg segmentation results along with the probability of normal ovaries. When masses are identified, the system isolates and normalizes each connected ROI component, then feeds the resulting image patches into OvaMTA-Diagnosis. This step enables detailed segmentation of the mass region and binary classification as benign or malignant. Ultimately, the system produces detection boxes outlining the segmented ovarian mass areas and their associated malignancy probabilities for each image, as depicted in **Figure 2**.

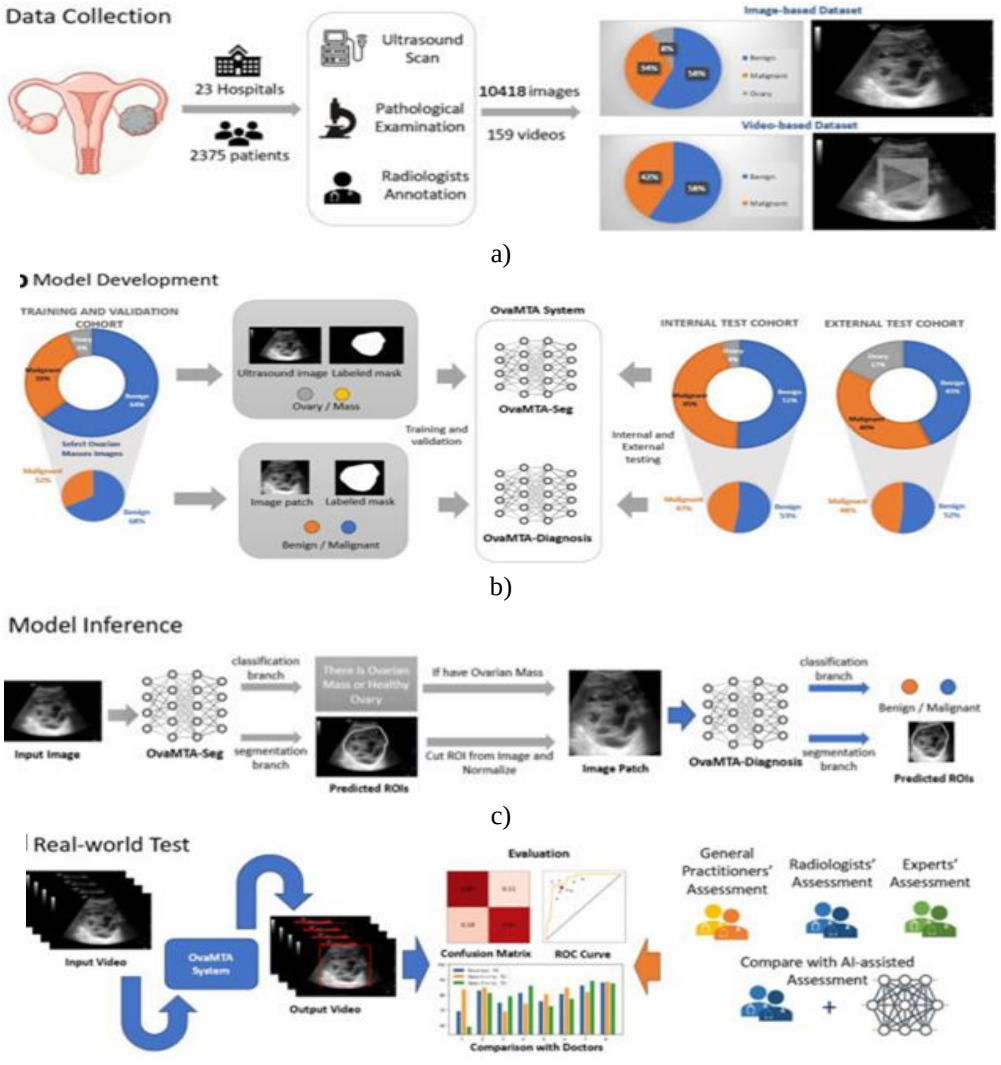


Figure 2. Schematic representation of the research framework.

(a) Ultrasound data and corresponding pathological outcomes were gathered from 23 hospitals, with boundaries of ovarian masses carefully annotated. (b) Creation of both an image dataset and a video dataset to support model construction. (c) The proposed OvaMTA platform integrates two neural networks: OvaMTA-Seg for automated tumor identification and OvaMTA-Diagnosis for malignancy prediction. (d) Model evaluation involved metrics such as ROC curves, confusion matrices, accuracy, sensitivity, specificity, and Kappa coefficient, alongside direct comparisons with physicians during video-based testing. ROC = receiver operating characteristic curve; OvaMTA = Ovarian Multi-Task Attention Network; ROI = region of interest; AI = artificial intelligence.

The two components of the system were trained independently. In line with their specific roles, the training data for OvaMTA-Seg encompassed ultrasound images from all training cases involving both ovaries and ovarian masses. By contrast, the training data for OvaMTA-Diagnosis was restricted to ultrasound images of ovarian masses from all training cases **(Figure 2)**. Both networks were constructed upon the MTANet architecture [27]. This is a single-stage multi-task attention framework featuring a Pyramid Vision Transformer (PVT) module for classification and a UNet-style multi-attention decoder for segmentation. The backbone was originally developed by the team and had previously surpassed leading models on the CVC ClinicDB dataset for polyp segmentation, the ISIC-2018 dataset for skin lesion segmentation, and a US dataset for liver tumor segmentation and classification [27]. It was therefore selected as the foundation for both OvaMTA-Seg and OvaMTA-Diagnosis.

Thresholds were established using a two-stage decision process informed by the output probability scores from OvaMTA on the training and validation data. If the probability of an ovarian mass in an image fell below 0.9, the image was classified as containing only healthy ovaries, and the system returned the ovary segmentation outcome. Conversely, when the probability exceeded 0.9, the image was deemed to contain ovarian masses. In such cases, connected regions from the OvaMTA-Seg output were isolated via bounding boxes, cropped, and passed to OvaMTA-Diagnosis for refined mass segmentation and benign/malignant determination. For OvaMTA-Diagnosis, patches with a malignancy probability above 0.37 were labeled malignant; otherwise, they were considered benign. It is important to note that these thresholds were determined empirically. The ideal cutoff for the same model may vary across different patient groups [29]. Here, the selected threshold corresponds to the point that maximizes the difference between the true positive rate and false positive rate, thereby helping to balance sensitivity and specificity to a reasonable degree [30].

Both OvaMTA-Diagnosis and OvaMTA-Seg were developed in PyTorch and trained on an NVIDIA GeForce RTX 3080 graphics processing unit. Training used a batch size of 10 and the AdamW optimizer with a learning rate of 1e-5 across all models. Each model was trained for 200 epochs with a learning rate scheduler that applied a 0.1 decay every 30 epochs. A randomly chosen validation subset, representing 10% of the combined training and validation data, was used to fine-tune the network parameters and hyperparameters. In the end, OvaMTA generates bounding boxes for the ROIs along with probability scores for healthy ovary, benign tumor, or malignant tumor. These three probabilities always sum to 1 and remain between 0 and 1. Segmentation performance was measured primarily by the average Dice coefficient. Classification performance was evaluated using multiple indicators, including sensitivity (SEN), specificity (SPE), accuracy (ACC), Cohen's kappa score, and the area under the receiver operating characteristic curve (AUC).

The complete implementation code for the deep learning system is publicly accessible at: https://github.com/SLYXDWL/OvaMTA.

Generalizing the model to real-world ultrasound diagnostic scenarios

For patient-level predictions based on images, the average of the predicted probabilities across multiple images from the same individual was calculated to produce a single overall score. For video-based predictions, each video frame was normalized individually and fed into OvaMTA. The system provided real-time diagnostic outputs for the current frame. The final benign or malignant classification for the entire video was derived from these frame-level results. To incorporate temporal context, feature maps from the preceding six frames were averaged to form the feature representation for the current frame. Binarized segmentation masks were produced using a 0.5 threshold. The resulting segmentation area for the current frame was expanded by 25 pixels to extract a patch, from which probabilities for healthy, benign, and malignant categories were computed for that frame. After processing all frames, the frame-level malignancy probabilities were averaged to determine the overall benign or malignant status of the video.

Comparison with radiologists and performance enhancement by AI

Six radiologists and two general practitioners participated in the evaluation of ovarian masses. This group included two junior radiologists (with at least 3 years of clinical experience), two intermediate radiologists (at least 5 years), two senior radiologists (at least 10 years), and two non-radiology practitioners (at least 3 years). The eight clinicians independently reviewed all anonymized and randomly ordered videos from external dataset B at the patient level, assigning a benign or malignant classification to each mass. Two weeks later, they re-evaluated the same videos with the assistance of OvaMTA outputs. The tumor boundaries segmented by OvaMTA, along with the bounding box and associated malignancy probability, were displayed alongside the original ultrasound video **(Figure 5)**. Video-level diagnostic conclusions from OvaMTA were also shown to the physicians at the conclusion of each video. All evaluators remained blinded to the final pathological diagnoses.

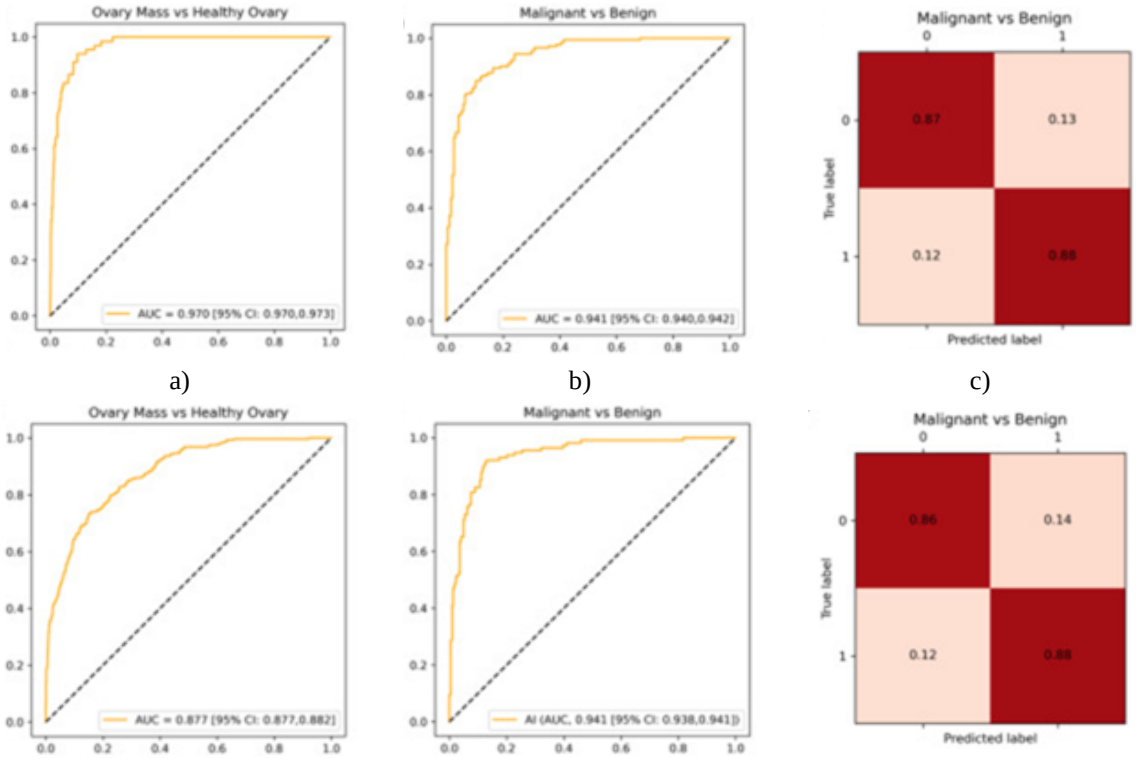
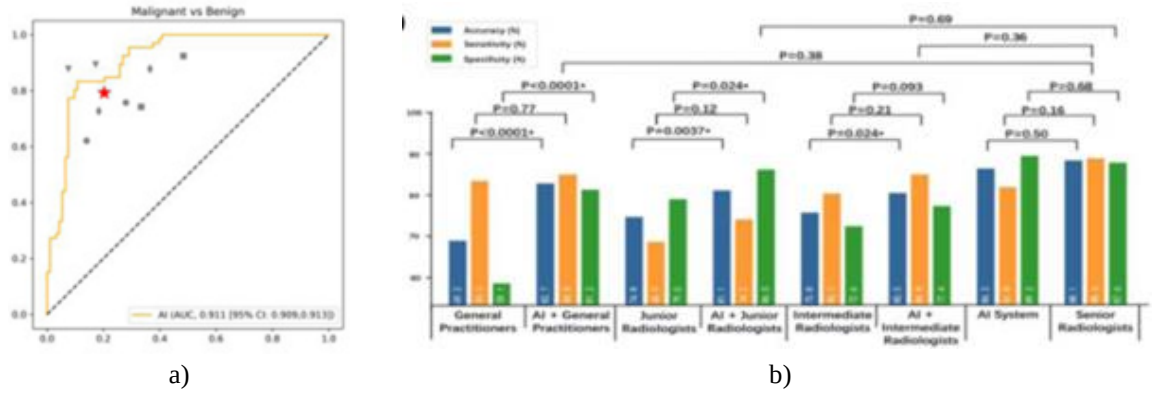
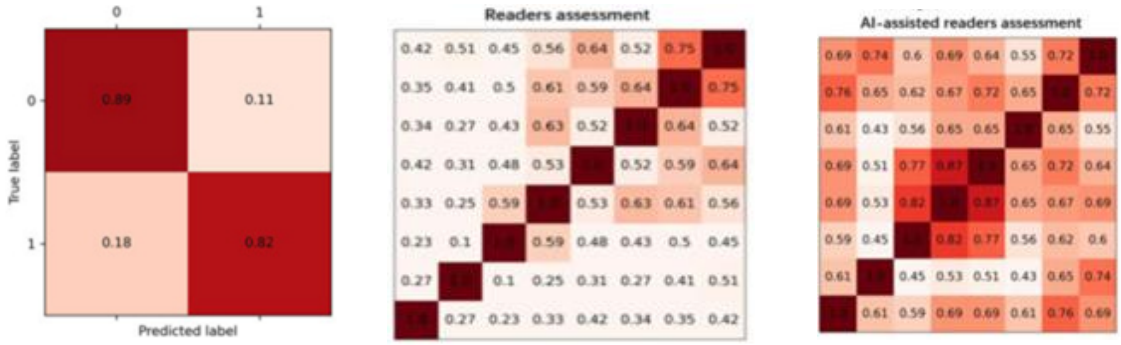


Figure 3. Diagnostic accuracy of OvaMTA during image-based evaluations.

Receiver operating characteristic (ROC) curves and confusion matrices for the internal test set (a, b, c) and external test set A (d, e, f) demonstrate the AI system's effectiveness in identifying ovarian masses and distinguishing between benign and malignant ovarian tumors using static images. ROC = receiver operating characteristic curve; OvaMTA = Ovarian Multi-Task Attention Network; AI = artificial intelligence.

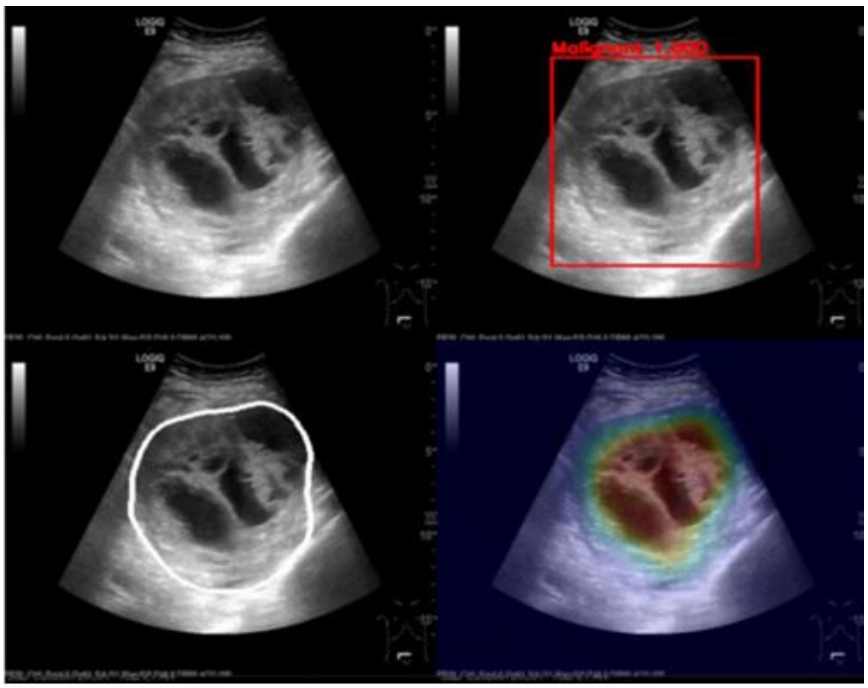




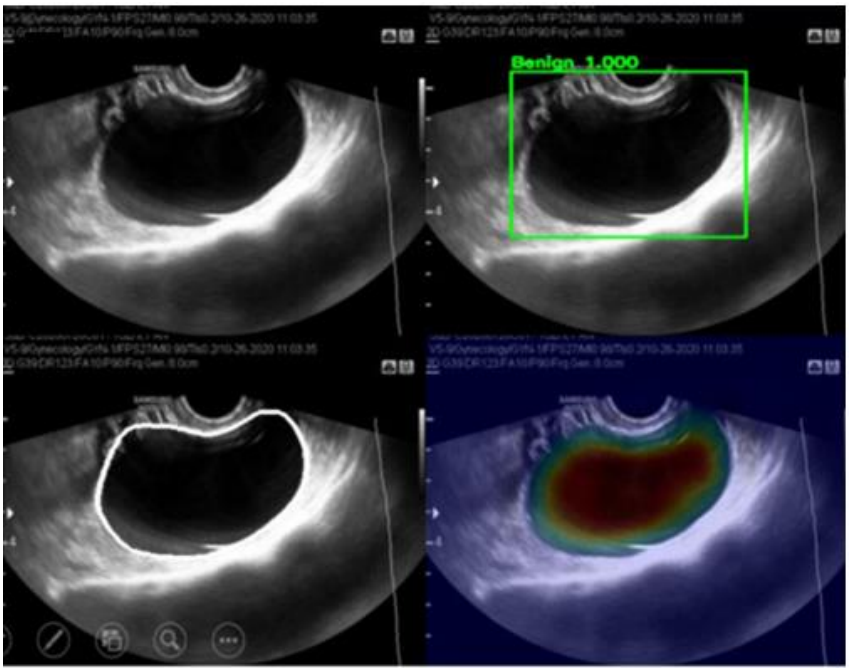
c) d) e)

Figure 4. Side-by-side evaluation of the diagnostic effectiveness of the AI platform versus radiologists with differing levels of expertise and general practitioners.

(a) Receiver operating characteristic (ROC) curves from the external video test set highlight the AI platform's classification ability, presented together with results from radiologists across experience categories. A prominent red star marks the overall average performance of the radiologists. Grey circles indicate outcomes for junior radiologists, grey diamonds for intermediate-level radiologists, grey triangles for senior experts, and grey squares for general practitioners. (b) Bar charts summarize the results of the eight participating clinicians, including p-values from paired McNemar tests to assess differences between groups. (c) Normalized confusion matrix detailing how well the AI platform separates benign from malignant ovarian tumors across the external video tests. (d) Kappa agreement matrix illustrating consistency among the eight doctors' unaided interpretations of the external test videos. (e) Kappa agreement matrix illustrating consistency among the eight doctors' interpretations when supported by the AI system on the external test videos. Every entry within the kappa matrices quantifies the pairwise agreement level between individual evaluators. On the axes of the kappa matrices in panels (d) and (e), general practitioners 1–2 and radiologists 1–6 are positioned from bottom to top vertically and left to right horizontally. ROC = receiver operating characteristic curve; AI = artificial intelligence.



a)



b)

Figure 5. Examples of two different ovarian masses identified, outlined, classified, and visually explained by the AI platform.

(a) Representative frame from a grayscale ultrasound video of a 64-year-old woman showing a multilocular cyst containing solid elements. The AI platform successfully locates the mass, delineates its boundaries, and delivers a real-time malignant classification along with the associated probability. The system's final output aligned with the postoperative pathological confirmation (endometrioid carcinoma). Real-time interpretable heat maps generated by the AI direct clinicians' focus toward irregular septations and solid areas. (b) Representative frame from a grayscale ultrasound video of a 52-year-old woman displaying a unilocular anechoic cyst. The AI platform identifies the mass, traces its contours, and provides a real-time benign classification with probability. This prediction matched the surgical pathology report (serous cystadenofibroma). The AI platform produces real-time interpretable heat maps that highlight the anechoic cystic regions for closer clinical attention. AI = artificial intelligence.

Statistics

Model development, which covered the training, validation, and internal testing phases, relied on a retrospective review that combined histopathological findings with follow-up records. Study participants were divided into training, validation, and internal testing groups at a 7:1:2 ratio. This resulted in 1514 individuals in the combined training and validation cohort and 363 individuals in the internal testing cohort. For performance assessment across the two external testing groups, the same inclusion and exclusion standards were applied, leading to 339 participants in external testing set A and 159 participants in external testing set B. The 95% confidence intervals (CI) for sensitivity, specificity, accuracy, Cohen's kappa score, and AUC of both the OvaMTA platform and the participating physicians were calculated via bootstrapping methods. McNemar's test for paired categorical data was employed to compare the sensitivity, specificity, and accuracy of the AI platform against those of the physicians and the AI-supported physicians. A p-value below 0.05 was considered statistically significant. All statistical computations were carried out in Python (version 3.10.12), utilizing the Scikit-learn (version 1.4.1.post1) and SciPy (version 1.11.2) libraries.

Role of funding source

The funding organizations had no involvement in the study design or execution; data collection, handling, analysis, or interpretation; manuscript drafting, revision, or approval; or the choice to publish the manuscript. W. Dai and Y. Wu had full access to the study data and took ultimate responsibility for the decision to submit the manuscript for publication.

Results and Discussion

Demography of the participants

This investigation incorporated data from 2375 patients in total (10,418 images and 159 videos) **(Figure 1)**. Information gathered from the 21 hospitals was partitioned for model building as follows: training and validation set (6,938 images from 1,514 patients) and internal test set (1,584 images from 363 patients). Data from patients at the two additional hospitals (1,896 images from 339 patients) formed external test set A. Key demographic and clinical features of the study population are summarized in **Table 1**. Age details were available for every participant. In the training and validation cohorts, patient ages spanned 12–86 years (median = 41). In the internal test cohort, ages ranged from 15–81 years (median = 45). For the external image and video test cohorts, ages ranged from 11–75 years (median = 39) and 17–81 years (median = 43), respectively.

Table 1. Baseline characteristics of patients.

Characteristic	Training and validation dataset (image)	Internal testing dataset (image)	External testing dataset A (image)	External testing dataset B (video)
No. of patients	1514	363	339	159
No. of samples	6938	1584	1896	159
Age (y)	42.05 ± 13.81 (12–86)	44.19 ± 14.91 (15–81)	40.35 ± 14.99 (11–75)	43.40 ± 14.65 (17–81)
No. of healthy ovary samples	431 (6.2)	67 (4.2)	322 (16.9)	–
No. of benign samples	4408 (63.5)	803 (50.7)	822 (43.3)	93
No. of malignant samples	2099 (30.3)	714 (45.1)	752 (39.6)	66
Tumor type				
Benign patients	966 (63.8)	183 (50.4)	224 (66.1)	93 (58.5)
Malignant patients	548 (36.1)	180 (49.6)	115 (33.9)	66 (41.5)
Histology				
Clear Cell	29 (1.9)	10 (2.7)	4 (1.2)	4 (2.5)
Endometrioid	35 (2.3)	10 (2.7)	6 (1.8)	6 (3.8)
Germ cell	293 (19.4)	63 (17.4)	146 (43.1)	15 (9.4)
Mucinous	103 (6.8)	45 (12.4)	36 (10.6)	31 (19.5)
Mixed[a]	73 (4.8)	13 (3.6)	20 (5.9)	6 (3.8)
Serous	375 (24.8)	128 (35.3)	67 (19.7)	58 (36.5)
Sex cord stromal tumor	61 (4.0)	24 (6.6)	19 (5.6)	11 (6.9)
Metastatic ovarian tumor	15 (1.0)	9 (2.5)	4 (1.2)	7 (4.4)
Others[b]	530 (35.0)	61 (16.8)	37 (10.9)	21 (13.2)

Note. The investigation involved 2375 patients in total (10,418 images and 159 videos). Average values are presented as mean ± SD, with the range shown in parentheses. Categorical data are reported as counts, with percentages included in parentheses.
SD = standard deviation.
[a] Mixed histology of ovarian mass refers to cases where the lesion simultaneously exhibits multiple pathological subtypes.
[b] Other histology categories included mesenchymal tumors, Brenner tumors, and similar entities. The samples comprise both images and videos.

Performance of AI system in ovary and ovarian mass segmentation

Segmentation quality was measured via the Dice coefficient across both the internal and external image test collections. OvaMTA-Seg recorded a mean Dice score of 0.887 ± 0.134 for the internal test group and 0.819 ± 0.201 for external test group A. Benign lesions showed Dice values of 0.912 ± 0.103 internally and 0.843 ± 0.192 externally. Malignant tumors achieved 0.878 ± 0.131 in the internal set and 0.851 ± 0.141 in the external set.

Figure 5 illustrates typical segmentation outcomes generated by OvaMTA-Seg on ultrasound scans. The AI framework supports highly accurate boundary definition in these images. In dynamic video evaluations, the system managed continuous identification and monitoring of masses throughout their visibility in the scan sequence. Consequently, this segmentation module can function independently as a helpful visualization resource for radiologists, effectively emphasizing suspicious regions.

Performance of AI system in ovarian mass detection

Given OvaMTA's stepwise decision approach, separate assessments were carried out for mass identification and for benign/malignant categorization on the internal and external image test collections. For detecting the existence of masses, OvaMTA-Seg produced an AUC of 0.970 (95% CI: 0.970–0.973) internally and 0.877 (95% CI: 0.877–0.882) in external test set A. Using an Intersection over Union (IoU) threshold above 0.5 to define correct detection, success rates stood at 94.2% internally and 88.1% on external images. In external test set B video assessments, the model reached a complete 100% detection rate, identifying every mass that appeared in the recordings.

In the task of spotting malignant lesions, the model delivered an AUC of 0.941 (95% CI: 0.940–0.942) on the internal test set and 0.941 (95% CI: 0.938–0.941) on external test set A **(Figure 3)**. For the video evaluations in external test set B, it attained an AUC of 0.911 (95% CI: 0.909–0.913) **(Figure 4)** together with 86.2% accuracy (137/159; 95% CI: 80.5–91.2) when classifying lesions as benign or malignant.

Comparison of diagnostic performance between AI system and doctors

Table 2 presents the outcomes from external test set B. Senior radiologists attained 88.1% accuracy (95% CI: 84.3–91.5), 88.6% sensitivity (95% CI: 82.9–93.9), and 87.6% specificity (95% CI: 82.6–92.3). Junior radiologists achieved 74.8% accuracy (95% CI: 70.1–79.6), 68.9% sensitivity (95% CI: 61.0–76.9), and 79.0% specificity (95% CI: 73.0–84.7). Intermediate radiologists recorded 75.8% accuracy (95% CI: 71.1–80.5), 80.3% sensitivity (95% CI: 73.1–86.9), and 72.6% specificity (95% CI: 66.3–79.2). The AI platform matched senior radiologists closely, posting 86.2% accuracy (137/159; 95% CI: 80.5–91.2; p = 0.50), 81.8% sensitivity (54/66; 95% CI: 72.3–91.0; p = 0.16), and 89.2% specificity (83/93; 95% CI: 82.8–95.2; p = 0.68). It significantly exceeded the results of intermediate radiologists (p = 0.00026), junior radiologists (p < 0.0001), and general practitioners (p < 0.0001). The AI system also showed better sensitivity than junior radiologists (p = 0.0046) and higher specificity than intermediate radiologists (p < 0.0001), junior radiologists (p = 0.0019), and general practitioners (p < 0.0001). Taken together, the AI platform's diagnostic strength clearly outperformed lower- and mid-level experienced radiologists while equaling the proficiency of senior-level experts.

Table 2. Diagnostic performances of AI system and doctors for detecting malignant ovarian tumors from ovarian masses.

Modes	Accuracy[a] (%)	*p* value	Sensitivity[a] (%)	*p* value	Specificity[a] (%)	*p* value	Kappa score
Internal test set (image-based)							
OvaMTA	87.3 (317/363) [83.7, 90.6]		87.8 (158/180) [82.7, 92.4]		86.9 (159/183) [81.8, 91.6]		0.747 [0.674, 0.813]
External test set A (image-based)							
OvaMTA	86.7 (294/339) [83.1, 90.2]		87.8 (101/115) [81.5, 93.4]		86.2 (193/224) [81.7, 90.5]		0.714 [0.636, 0.791]
External test set B (video-based)							
OvaMTA	86.2 (137/159) [80.5, 91.2]		81.8 (54/66) [72.3, 91.0]		89.2 (83/93) [82.8, 95.2]		0.714 [0.601, 0.817]
Senior radiologists	88.1 [84.3, 91.5]	0.50	88.6 [82.9, 93.9]	0.16	87.6 [82.6, 92.3]	0.68	0.756 [0.678, 0.825]
Intermediate Radiologists	75.8 [71.1, 80.5]∗	0.00026	80.3 [73.1, 86.9]	0.87	72.6 [66.3, 79.2]∗	<0.0001	0.515 [0.418, 0.606]
AI-assisted intermediate radiologists	80.5 [76.1, 84.9]∗	0.018	84.8 [78.4, 90.8]	0.54	77.4 [71.4, 83.4]∗	<0.0001	0.608 [0.519, 0.693]
Junior radiologists	74.8 [70.1, 79.6]∗	<0.0001	68.9 [61.0, 76.9]∗	0.0046	79.0 [73.0, 84.7]∗	0.0019	0.481 [0.382, 0.578]
AI-assisted junior radiologists	81.1 [76.7, 85.5]∗	0.0025	74.2 [66.7, 81.7]∗	0.031	86.0 [80.9, 90.9]	0.070	0.608 [0.517, 0.697]

General practitioners	69.2 [64.2, 74.2]∗	<0.0001	83.3 [76.8, 89.6]	0.88	59.1 [52.1, 66.0]∗	<0.0001	0.401 [0.308, 0.493]
AI-assisted general practitioners	82.7 [78.3, 86.8]	0.19	84.8 [78.5, 90.7]	0.60	81.2 [75.4, 86.7]∗	0.0041	0.650 [0.565, 0.732]

Note. Values shown in parentheses indicate numerators/denominators; values in brackets represent 95% confidence intervals.
∗$p < 0.05$, denoting a statistically significant difference compared with the AI system.
AI = artificial intelligence. OvaMTA = Ovarian Multi-Task Attention Network. CI = confidence interval.
[a] Comparisons of differences between the clinicians and the AI system were performed, with corresponding p-values calculated.

Auxiliary diagnosis function of AI system on ultrasound videos

Support from OvaMTA led to a marked rise in general practitioners' accuracy, increasing from 69.2% to 82.7% ($p < 0.0001$). Junior radiologists saw their accuracy improve from 74.8% to 81.1% ($p = 0.0037$), while intermediate radiologists improved from 75.8% to 80.5% ($p = 0.024$). Overall, AI assistance produced substantial gains for junior and intermediate radiologists relative to their unaided results (accuracy: 80.8% vs. 75.3%, $p = 0.00015$; sensitivity: 79.5% vs. 74.6%, $p = 0.029$; specificity: 81.7% vs. 75.8%, $p = 0.0032$). The AI platform notably boosted specificity among less experienced clinicians, elevating general practitioners from 59.1% to 81.2% ($p < 0.0001$) and junior radiologists from 79.0% to 86.0% ($p = 0.024$). With AI support, junior radiologists' specificity reached a level statistically equivalent to senior radiologists (86.0% vs. 87.6%, $p = 0.69$). Similarly, the sensitivity of non-radiology practitioners with AI assistance matched that of senior radiologists (84.8% vs. 88.6%, $p = 0.38$). Assisted general practitioners attained overall performance comparable to the average of radiologists (accuracy: 82.7% vs. 81.8%, $p = 0.80$; sensitivity: 84.8% vs. 82.6%, $p = 0.72$; specificity: 81.2% vs. 81.2%, $p > 0.99$).

The Kappa matrices displayed in **Figure 4** (d) and (e) indicate that AI assistance increased consistency in diagnoses across different clinicians. Thus, the AI system not only raised individual diagnostic proficiency but also enhanced agreement between observers.

Visualization and interpretability of AI system

As illustrated in **Figure 5** and the accompanying videos, OvaMTA accurately identifies tumor locations, delineates adjacent ovarian tissue, and conducts frame-by-frame analysis throughout the video. The model supports clinicians in determining whether ovarian masses are benign or malignant. Videos S1 and S2 demonstrate the AI platform's stable, continuous production of segmentation and diagnostic outputs for each frame.

Figure 5 (a) displays a representative frame from a grayscale ultrasound video (Video S1) of a 64-year-old woman presenting with solid components inside a multilocular cyst. The AI platform detects the mass, outlines its margins, and supplies real-time malignant classification along with probability scores. Its final assessment agreed with the postsurgical pathology (endometrioid carcinoma). Real-time interpretable heat maps from the AI help direct clinicians' attention to irregular divisions and solid portions. **Figure 5** (b) shows a representative frame from a grayscale ultrasound video (Video S2) of a 52-year-old woman with a unilocular anechoic cyst. The AI platform identifies the lesion, traces its edges, and issues real-time benign classification and probability values. This matched the surgical pathology (serous cystadenofibroma). The AI system generates real-time interpretable heat maps that highlight the anechoic cystic areas for focused review.

Precise detection and diagnosis of ovarian cancer play a vital role in enhancing patient survival and long-term outcomes. Recent progress in deep learning has shown its potential for evaluating ovarian cancer through ultrasound (US) imaging. Nevertheless, no existing AI pipeline fully aligns with routine clinical scanning procedures. To address this, we developed OvaMTA, an automated deep learning pipeline for ovarian mass identification and pathological classification based on US screening. The system was rigorously evaluated using both static images and dynamic videos. It demonstrated strong capabilities in localizing ovarian masses and normal ovaries in ultrasound scans (mean Dice: 0.819). Additionally, it exhibited excellent performance in distinguishing benign from malignant masses in ultrasound videos (AUC: 0.911). The ultrasound-based OvaMTA platform offers a straightforward and effective support tool for ovarian cancer screening, achieving diagnostic accuracy similar to senior radiologists. It therefore represents a promising resource for ovarian cancer detection and diagnosis.

OvaMTA functions as a comprehensive platform integrating detection, segmentation, and diagnosis of ovaries and associated masses. During ovarian cancer screening scans, locating regions of interest in US images is critical,

as radiologists require complete measurements and detailed observation to assess ovarian or mass status. Earlier research primarily addressed ovarian masses alone [10, 12, 14-17, 18, 31, 32], often relying on restricted datasets or omitting robust external validation. Many models were designed for post-scan analysis after image collection and clinical data review, limiting their utility in fast-paced clinical environments. Developing a fully automated segmentation module tailored to ovarian ultrasound is therefore essential. Such a component not only boosts overall AI precision but also partially improves model transparency.

OvaMTA delivered state-of-the-art results. The study benefited from multicenter data, inclusion of healthy ovarian annotations, and an advanced network design that optimized clinical workflow. The underlying MTANet backbone surpassed commonly used architectures such as ResNet and DenseNet [33, 34] in ovarian cancer imaging tasks [16, 18], as it seamlessly merges segmentation and classification, allowing mutual enhancement between the tasks and yielding superior accuracy [27]. Diverse training data further enabled better adaptation across varied patient populations. The system achieved higher mean Intersection over Union scores than prior works [35, 36] and outperformed the pooled accuracy reported in a recent meta-analysis (sensitivity: 0.90 vs. 0.88; specificity: 0.86 vs. 0.84) [37].

OvaMTA exhibited robust generalization, stability, and strong potential for clinical integration. Multicenter external validation combined with video-level testing offered more thorough assessment of real-world applicability than many previous studies. The fully automated pipeline supports seamless incorporation into physicians' daily workflow, moving beyond post-diagnosis verification. Given the deep pelvic location of ovarian structures, identifying and evaluating ovaries and masses can be difficult. The AI platform aids clinicians in detection, followed by detailed segmentation and classification with accompanying probability scores that provide intuitive guidance. For better explainability, real-time attention heat maps were generated from current-frame features, helping users understand model focus areas (Videos S1 and S2). The system substantially raised junior radiologists' accuracy in benign/malignant classification during screening (from 74.8% to 81.1%) and can support real-time ultrasound examinations. Furthermore, with brief training, it enables non-specialist general practitioners to perform at the average level of radiologists (accuracy: 82.7% vs. 81.8%, $p = 0.80$; sensitivity: 84.8% vs. 82.6%, $p = 0.72$; specificity: 81.2% vs. 81.2%, $p > 0.99$). This expands its usefulness to settings such as screening programs in primary care facilities.

Several limitations should be acknowledged. First, regional differences observed in external testing suggest possible threshold shifts. The thresholds reported here serve as effective reference points, but applying domain generalization methods could further refine the model [38]. Second, training videos included only cases with ovarian masses; thus, segmentation of normal ovaries was evaluated solely on static images rather than videos. Recognizing normal ovaries is particularly important in postmenopausal screening. The study also excluded other pelvic masses and scenarios involving multiple targets (e.g., coexisting masses and normal ovaries). Future screening-oriented models would benefit from expanded training with negative and normal samples from healthy individuals [39]. Third, suboptimal segmentation occasionally reduced diagnostic accuracy, indicating opportunities to optimize human-AI collaboration [40, 41]. Fourth, the current pipeline lacks an automated keyframe selection mechanism, which can introduce bias in video-level conclusions when non-representative frames influence overall predictions. Dense frame-by-frame labeling was not feasible due to misalignment with typical clinical annotation practices. Developing decision algorithms that better match physicians' scanning patterns is therefore warranted. Finally, color Doppler and clinical staging were not incorporated, as interpretations of Doppler findings vary among clinicians and may introduce inconsistency. Future work will examine clinical staging and multimodal approaches to refine ultrasound tools for ovarian masses.

Conclusion

In conclusion, the image-based OvaMTA platform effectively identifies and classifies normal ovaries and ovarian masses while differentiating benign from malignant lesions, with performance comparable to expert subjective evaluation. The AI system markedly enhances diagnostic capabilities for junior and intermediate radiologists. General practitioners aided by AI can reach the average performance level of radiologists. OvaMTA holds considerable promise as a practical, broadly applicable aid for physicians in ovarian mass screening and ovarian cancer diagnosis.

Acknowledgments: None

Conflict of Interest: None

Financial Support: This research received funding from the National Natural Science Foundation of China (Grant Nos. 12090020, 82071929, and 12090025) and the R&D project of the Pazhou Lab (Huangpu) (Grant No. 2023K0605).

Ethics Statement: None

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63.
2. Zheng R, Zhang S, Zeng H, Wang S, Sun K, Chen R, et al. Cancer incidence and mortality in China, 2016. J Natl Cancer Cent. 2022;2(1):1-9.
3. Lheureux S, Braunstein M, Oza AM. Epithelial ovarian cancer: evolution of management in the era of precision medicine. CA Cancer J Clin. 2019;69(4):280-304.
4. Froyman W, Landolfo C, De Cock B, Wynants L, Sladkevicius P, Testa AC, et al. Risk of complications in patients with conservatively managed ovarian tumours (IOTA5): a 2-year interim analysis of a multicentre, prospective, cohort study. Lancet Oncol. 2019;20(3):448-58.
5. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. Br J Obstet Gynaecol. 1990;97(10):922-9.
6. Amor F, Vaccaro H, Alcázar JL, León M, Craig JM, Martinez J. Gynecologic imaging reporting and data system: a new proposal for classifying adnexal masses on the basis of sonographic findings. J Ultrasound Med. 2009;28(3):285-91.
7. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol. 2008;31(6):681-90.
8. Van Calster B, Van Hoorde K, Valentin L, Testa AC, Froyman W, Wynants L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. BMJ. 2014;349:g5920.
9. Andreotti RF, Timmerman D, Strachowski LM, Froyman W, Benacerraf BR, Bennett GL, et al. O-RADS US risk stratification and management system: a consensus guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee. Radiology. 2020;294(1):168-85.
10. Kongara SR, Prakasha S, Brindha A. Performance evaluation of optimized convolutional neural network mechanism in the detection and classification of ovarian cancer. Multimed Tool Appl. 2024. doi:10.1007/s11042-024-18115-0
11. Xiang H, Xiao Y, Li F, Zhang H, Chen L, Wang Y, et al. Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. Nat Commun. 2024;15(1):2681.
12. Wang Z, Luo S, Chen J, Li Q, Zhang H, Liu Y, et al. Multi-modality deep learning model reaches high prediction accuracy in the diagnosis of ovarian cancer. iScience. 2024;27(4):109403.
13. Miao K, Zhao N, Lv Q, Wang Y, Zhang H, Li F, et al. Prediction of benign and malignant ovarian tumors using Resnet 34 on ultrasound images. J Obstet Gynaecol Res. 2023;49(12):2910-7.
14. Jung Y, Kim T, Han MR, Kim S, Lee J, Park H, et al. Ovarian tumor diagnosis using deep convolutional neural networks and a denoising convolutional autoencoder. Sci Rep. 2022;12(1):17024.
15. Hsu ST, Su YJ, Hung CH, Chen MJ, Lu CH, Kuo CE. Automatic ovarian tumors recognition system based on ensemble convolutional neural network with ultrasound imaging. BMC Med Inform Decis Mak. 2022;22(1):298.
16. Gao Y, Zeng S, Xu X, Zhang H, Chen L, Wang Y, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. Lancet Digit Health. 2022;4(3):e179-e187.

17. Christiansen F, Epstein EL, Smedberg E, Akerlund M, Smith K, Epstein E. Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment. Ultrasound Obstet Gynecol. 2021;57(1):155.
18. Chen H, Yang BW, Qian L, Wang Y, Zhang H, Chen L, et al. Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. Radiology. 2022;304(1):106-13.
19. Fan JF, Liu JQ, Chen QL, Wang W, Wu YH. Accurate ovarian cyst classification with a lightweight deep learning model for ultrasound images. IEEE Access. 2023;11:110681-91.
20. C K, Shivaram JM. Segmentation of ovarian cyst using improved U-NET and hybrid deep learning model. Multimed Tool Appl. 2023;83(14):42645-79.
21. Li J, Chen Y, Zhang M, Wang L, Liu S, Zhao H, et al. A deep learning model system for diagnosis and management of adnexal masses. Cancers. 2022;14(21).
22. Moro F, Vagni M, Tran HE, Mascilini F, Pignata S, Scambia G, et al. Radiomics analysis of ultrasound images to discriminate between benign and malignant adnexal masses with solid ultrasound morphology. Ultrasound Obstet Gynecol. 2024. doi:10.1002/uog.27680
23. Fanizzi A, Arezzo F, Cormio G, Loiudice L, Saponaro C, Mencattini A, et al. An explainable machine learning model to solid adnexal masses diagnosis based on clinical data and qualitative ultrasound indicators. Cancer Med. 2024;13(12):e7425.
24. Barcroft JF, Linton-Reid K, Landolfo C, Al-Memar M, Saso S, Stalder C, et al. Machine learning and radiomics for segmentation and classification of adnexal masses on ultrasound. NPJ Precis Oncol. 2024;8(1):41.
25. Li H, Fang J, Liu S, Zhang Y, Wang Z, Chen L, et al. CR-unet: a composite network for ovary and follicle segmentation in ultrasound images. IEEE J Biomed Health Inform. 2020;24(4):974-83.
26. Herrington CS. WHO classification of tumours female genital tumours. Lyon: International Agency for Research on Cancer; 2020.
27. Ling Y, Wang Y, Dai W, Yu J, Liang P, Kong D. MTANet: multi-task attention network for automatic medical image segmentation and classification. IEEE Trans Med Imag. 2024;43(2):674-85.
28. Kim TLD, Park EK, Choi S. Deep learning techniques of fatty liver using multiview ultrasound images scanned by different scanners. JMIR Med Inform. 2021;9(11):e30066.
29. Huang X, Wang Z, Zhang M, Luo H. Diagnostic accuracy of the ADNEX model for ovarian cancer at the 15% cut-off value: a systematic review and meta-analysis. Front Oncol. 2021;11:684257.
30. Chaurasia AK, MacGregor S, Craig JE, Mackey DA, Hewitt AW. Assessing the efficacy of synthetic optic disc images for detecting glaucomatous optic neuropathy using deep learning. Transl Vis Sci Technol. 2024;13(6):1.
31. Srivastava S, Kumar P, Chaudhry V, Singh A. Detection of ovarian cyst in ultrasound images using fine-tuned VGG-16 deep learning network. SN Comput Sci. 2020;1(2).
32. Barcroft JF, Linton-Reid K, Landolfo C, Al-Memar M, Saso S, Stalder C, et al. The use of machine learning models and radiomics for segmentation and classification of adnexal masses on ultrasound: a multi-cohort retrospective study. 2023.
33. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 2261-9.
34. He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770-8.
35. Zhao SL, Bai W, Cai L. MMOTU: a multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. arXiv. 2023. doi:10.48550/arXiv.2207.06799
36. Shantharam R, Palanisamy R. Analysis of hyperparameter tuned UNet++ deep model for delineation of ultrasound ovarian tumors. In: 2023 IEEE 7th Conference on Information and Communication Technology (CICT); 2023. p. 1.
37. Xu HL, Gong TT, Liu FH, Chen HY, Xiao Q, Hou Y, et al. Artificial intelligence performance in image-based ovarian cancer identification: a systematic review and meta-analysis. eClinicalMedicine. 2022;53:101662.
38. Wang J, Lan C, Liu C, Ouyang Y, Qin T, Lu W, et al. Generalizing to unseen domains: a survey on domain generalization. IEEE Trans Knowl Data Eng. 2022;35(8):8052-72.

39. He Y, Guo Y, Lyu J, Zhang H, Chen L, Wang Y, et al. Disorder-free data are all you need — inverse supervised learning for broad-spectrum head disorder detection. 2024;1(4):AIoa2300137.
40. Xu W, Jia X, Mei Z, Zhang H, Chen L, Wang Y, et al. Generalizability and diagnostic performance of AI models for thyroid US. Radiology. 2023;307(5):e221157.
41. He B, Kwan AC, Cho JH, Lu Y, Zhang H, Chen L, et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. Nature. 2023;616(7957):520-4.