Development and validation of a deep learning pipeline to diagnose ovarian masses using ultrasound screening: a retrospective multicenter study

Wen-Li Dai,[a,g] Ying-Nan Wu,[b,g] Ya-Ting Ling,[a] Jing Zhao,[c] Shuang Zhang,[d] Zhao-Wen Gu,[e] Li-Ping Gong,[b] Man-Ning Zhu,[b] Shuang Dong,[b] Song-Cheng Xu,[b] Lei Wu,[f] Li-Tao Sun,[b,**] and De-Xing Kong[a,*]

[a]School of Mathematical Sciences, Zhejiang University, Zijingang Campus, Hangzhou, Zhejiang, China
[b]Cancer Center, Department of Ultrasound Medicine, Zhejiang Provincial People's Hospital (Affiliated People's Hospital), Hangzhou Medical College, Hangzhou, Zhejiang, China
[c]Department of Ultrasound Medicine, Sichuan Provincial Maternity and Child Health Care Hospital, Chengdu, Sichuan, China
[d]Department of Ultrasound Medicine, Second Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang, China
[e]Department of Neurosurgery, The Second Affiliated Hospital of Zhejiang University School of Medicine, 88, Jiefang Road, Hangzhou, China
[f]Department of Ultrasound Medicine, Chongqing University Fuling Hospital, Chongqing, China

eClinicalMedicine
2024;78: 102923
Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102923

Summary

Background Ovarian cancer has the highest mortality rate among gynaecological malignancies and is initially screened using ultrasound. Owing to the high complexity of ultrasound images of ovarian masses and the anatomical characteristics of the deep pelvic cavity, subjective assessment requires extensive experience and skill. Therefore, detecting the ovaries and ovarian masses and diagnose ovarian cancer are challenging. In the present study, we aimed to develop an automated deep learning framework, the Ovarian Multi-Task Attention Network (OvaMTA), for ovary and ovarian mass detection, segmentation, and classification, as well as further diagnosis of ovarian masses based on ultrasound screening.

Methods Between June 2020 and May 2022, the OvaMTA model was trained, validated and tested on a training and validation cohort including 6938 images and an internal testing cohort including 1584 images which were recruited from 21 hospitals involving women who underwent ultrasound examinations for ovarian masses. Subsequently, we recruited two external test cohorts from another two hospitals. We obtained 1896 images between February 2024 and April 2024 as image-based external test dataset, and further obtained 159 videos for the video-based external test dataset between April 2024 and May 2024. We developed an artificial intelligence (AI) system (termed OvaMTA) to diagnose ovarian masses using ultrasound screening. It includes two models: an entire image-based segmentation model, OvaMTA-Seg, for ovary detection and a diagnosis model, OvaMTA-Diagnosis, for predicting the pathological type of ovarian mass using image patches cropped by OvaMTA-Seg. The performance of the system was evaluated in one internal and two external validation cohorts, and compared with doctors' assessments in real-world testing. We recruited eight physicians to assess the real-world data. The value of the system in assisting doctors with diagnosis was also evaluated.

Findings In terms of segmentation, OvaMTA-Seg achieved an average Dice score of 0.887 on the internal test set and 0.819 on the image-based external test set. OvaMTA-Seg also performed well in ovarian mass detection from test images, including healthy ovaries and masses (internal test area under the curve [AUC]: 0.970; external test AUC: 0.877). In terms of classification diagnosis prediction, OvaMTA-Diagnosis demonstrated high performance on image-based internal (AUC: 0.941) and external test sets (AUC: 0.941). In video-based external testing, OvaMTA recognised 159 videos with ovarian masses with AUC of 0.911, and is comparable to the performance of senior radiologists (ACC: 86.2 vs. 88.1, $p = 0.50$; SEN: 81.8 vs. 88.6, $p = 0.16$; SPE: 89.2 vs. 87.6, $p = 0.68$). There was a significant improvement in junior and intermediate radiologists who were assisted by AI compared to those who were not assisted by AI (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 assisted by AI achieved an average performance of radiologists (ACC: 82.7 vs. 81.8, $p = 0.80$; SEN: 84.8 vs. 82.6, $p = 0.72$; SPE: 81.2 vs. 81.2, $p > 0.99$).

*Corresponding author.
**Corresponding author.
E-mail addresses: dxkong@zju.edu.cn (D.-X. Kong), sunlitao@hmc.edu.cn (L.-T. Sun).
[g]W.-L. Dai. and Y.-N. Wu. contributed equally to this work.

Interpretation The OvaMTA system based on ultrasound imaging is a simple and practical auxiliary tool for screening for ovarian cancer, with a diagnostic performance comparable to that of senior radiologists. This provides a potential tool for screening ovarian cancer.

Funding This work was supported by 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).

Copyright © 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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

Research in context

Evidence before this study

We searched for the term "ultrasound AND deep learning AND (ovarian mass OR ovarian cancer) AND (diagnosis OR detection OR screen)" on PubMed, Web of Science, IEEE Xplore and Dimensions to obtain papers published from database establishment until 9 April 2024 without language restrictions. Some studies have focused on the application of Convolutional Neural Networks for the classification of benign and malignant ovarian masses using ultrasound images. Most of these applications require manual cropping of ultrasound images which is not conducive to use in clinical settings. There is no research on deep-learning pipelines to detect, segment, and classify ovaries and ovarian masses, and further discriminate ovarian masses, which is based on multicenter and further validation at the video level, conforming to the clinical workflow. Therefore, it is necessary to develop an effective and convenient method to detect, locate, and diagnose the relevant areas during ultrasound scanning to fully leverage the role of deep learning in ovarian ultrasound.

Added value of this study

This paper proposes an automated artificial intelligence (AI) system called Ovarian Multi-Task Attention Network (OvaMTA) for tumour localisation, ovarian localisation, and pathological type diagnosis based on ultrasound scanning. To adapt to clinical scenarios, unlike previous studies that were limited to images, our model was designed for dynamic scanning videos, and we included healthy ovarian annotations for training. We compared OvaMTA with general practitioners and radiologists in external validation cohorts.

Implications of all the available evidence

Our system demonstrated promising performance in locating tumour and ovaries in ultrasound scanning videos. Additionally, it also possesses diagnostic capabilities for differentiating between benign and malignant ovarian masses that are on a par with those of experienced radiologists. This has the potential to benefit non-radiological doctors, junior radiologists, and intermediate radiologists, offering a practical and clinically useful aid in the detection and diagnosis of ovarian cancer.

Introduction

Ovarian cancer is one of the three most common malignant tumours in gynaecology.[1] With the highest mortality rate among gynaecological malignancies, and has become an important threat to women's lives and health.[2] Ovarian cancer occurs deep in the pelvis, with a hidden onset and no notable clinical symptoms, making it prone to misdiagnosis and missed diagnosis.[3] Ultrasound (US) imaging plays an important role in the detection of ovarian cancer. Due to the complexity of the anatomical location and biological behaviour of ovarian cancer, as well as the extremely time-consuming and operator-dependent nature of US imaging, the accuracy of US diagnosis of ovarian cancer alone is still relatively low.[4] Therefore, to improve the diagnostic accuracy of ovarian cancer, numerous studies have been conducted on various clinical diagnostic models and classification systems for ovarian masses based on general patient clinical data, US image features, and laboratory tests, such as 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] However, the current US diagnostic methods for ovarian masses rely mostly on manual feature extraction by professional radiologists, and their diagnostic accuracy and efficiency need to be improved. Without practicality and universality, they are not conducive to widespread promotion and application in grassroots medical institutions. Therefore, it is important to develop a rapid and accurate automatic diagnostic system for ovarian cancer.

Recent developments in deep learning (DL) algorithms have shown that US diagnosis can benefit significantly from its ability to automatically extract complex information from images without requiring radiologist-selected features. Several studies have explored DL algorithms for evaluating ovarian masses and acquired good performance.[10–24] However, a fully automated artificial intelligence (AI) pipeline that is

robust to varying US image conditions and aligned with the standards of actual clinical application is lacking. First, because of the anatomical characteristics of the ovaries and masses located deep in the pelvic cavity, it is challenging for physicians to detect and segment them ovaries and masses. Present studies largely depend on US images labelled with a priori knowledge assessment and segmentation.[19,25] However, preprocessing images manually is a time-consuming and highly professional process, and may introduce observer variability. To achieve full automation of DL applications, the integration of automatic segmentation and classification models is required. Second, in real clinical scenarios, doctors are based on real-time dynamic observations and the diagnosis of ovarian masses; however, existing studies are only applicable to the diagnosis of static images. Third, the existing studies lack real-time visualisation and interpretability.

We aimed to develop and validate a deep-learning pipeline Ovarian Multi-Task Attention Network (OvaMTA) for real-time navigation of diagnostic planes during US scanning to identify and label healthy ovaries and ovarian masses, and further screen for ovarian cancer using heterogeneous multicenter datasets of images for retrospective testing and external validation. The scalability of the tool was evaluated using a real-world clinical video dataset. The performances of junior radiologists and non-US general practitioners with AI-assistance were compared with those of junior, intermediate, and senior radiologists. We also attempt to visually interpret the "black box" in real-time of the system's internal decision-making. The AI system performed well in terms of tumour and ovarian localisation in the US scanning videos. At the same time, it has excellent diagnostic ability for benign and malignant ovarian masses and has improved the diagnostic accuracy of doctors, including general practitioners and radiologists. This method may provide a potential tool for ovarian cancer detection and diagnosis.

Methods

Ethics

This was a retrospective multicenter study that used US image sets obtained from 23 hospitals. We retrospectively collected grayscale US images from the Zhejiang Provincial People's Hospital and National Health Commission Medical Imaging Standard Database of China, including 21 hospitals, from June 2020 to May 2022 as training, validation, and internal test datasets. The study protocol was approved by the Institutional Review Board of the Zhejiang Provincial People's Hospital (QT2023117). The requirement for written informed consent was waived. We obtained images from the Sichuan Provincial Maternity and Child Health Care Hospital and the Second Affiliated Hospital of Harbin Medical University between February 2024 and April 2024 as external test dataset A. We further obtained videos for the external validation dataset B from the two centres between April 2024 and May 2024. The study protocols were approved by the Institutional Review Boards of the Sichuan Provincial Maternity and Child Health Care Hospital (20240205-011) and the Second Affiliated Hospital of Harbin Medical University (KY2024-029). Written informed consent was obtained from all participants before their enrolment in the study.

Inclusion and exclusion criteria

Inclusion criteria were as follows: patients with at least one persistent ovarian mass detected at US. For training, validation, and internal test datasets collected from June 2020 to May 2022, all patients underwent surgery following US examination and histopathological results were obtained (1636 cases), or had an US follow-up of at least 6 months until the lesion resolved (241 cases). Our subsequent case studies exclusively focused on follicular cysts and corpus luteum cysts, disregarding any consideration of other types within our patient cohort. For external test dataset A collected between February 2024 and April 2024 and external test dataset B collected between April 2024 and May 2024, all the patients' ovarian masses were confirmed through pathological examination (498 cases). During the US examination of patients with unilateral ovarian masses, images of ovaries on the opposite side without masses were classified as normal ovaries also included in the study. The pathological diagnosis was made in accordance with the WHO classification of female genital tumours (2020) and classified into benign and malignant tumours (including borderline tumours).[26] The ovarian masses in this study included ovarian and fallopian tube-origin tumours.

Exclusion criteria were as follows: patients with non-ovarian tumours other than metastatic ovarian tumours confirmed by histopathological analysis, those with uncertain histopathological results, and those without follow-up ultrasound results were excluded. When multiple ovarian masses were presents in one image, the image was excluded.

Data were collected retrospectively by trained investigators using a unified form and reviewed by two researchers (Y.N.W. and W.L.D). A flowchart of this process is shown in Fig. 1. The study was conducted in compliance with the STARD-2015 guidelines (equator-network.org).

Ultrasound examinations

All US examinations were performed using 38 types of equipment from 9 vendors with transvaginal, transrectal, and transabdominal scanning method (Supplementary Table S1). Preoperative ovarian ultrasonography images and videos were obtained by radiologists with >5 years of experience in pelvis ultrasonography, following the

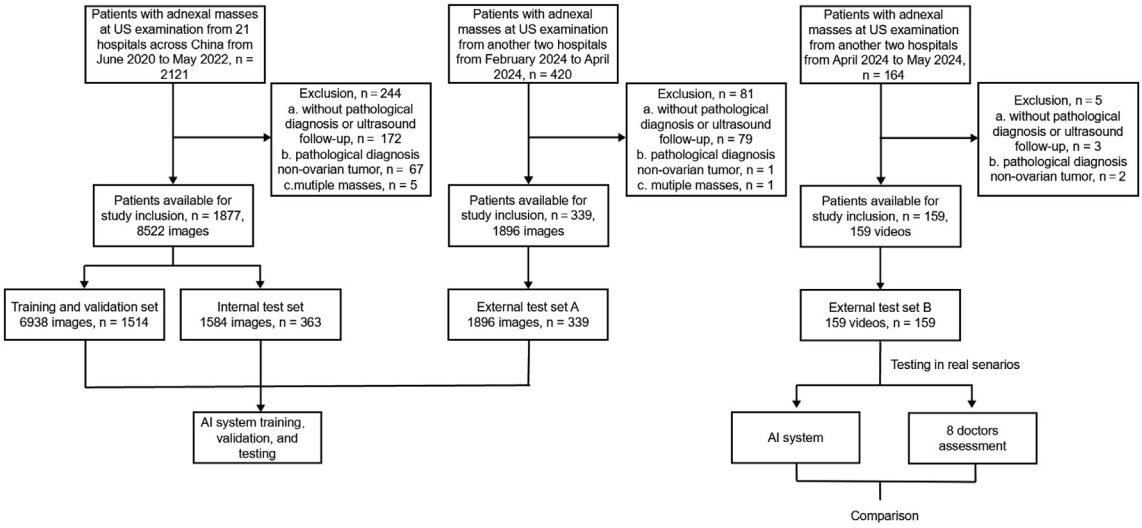

Fig. 1: Flowchart shows the eligibility criteria and procedure of AI system evaluation. US, ultrasound. AI, artificial intelligence.

standard protocol. A grayscale dynamic video of scanning was recorded for the ovarian masses at a uniform velocity for at least 2–5 s. The video was routinely recorded from "none" to "the maximum" to "none" during the scan. The region of interest (ROI) was positioned at the centre of the image to capture a standardised US video. The data stored and used for analysis included a longitudinal maximum cross-sectional US image and a maximum cross-sectional US image, as well as at least one typical imaging feature image, such as papillary structures, posterior acoustic shadows, and solid parts. Images of healthy ovarian tissues from patients were also stored and used for analysis, including at least one longitudinal ultrasound image and one transverse ultrasound image.

Training dataset preprocessing

All ovarian US images were converted to the PNG format. Next, a label system developed by Zhejiang Demetics Medical Technology Co., Ltd. with software version 4.4.0.1 was used to outline all ovarian masses and healthy ovaries. The work of labelling and sketching the ROI in the above pictures was completed by a radiologist with 8 years of experience in US, which was then reviewed and modified by a radiologist with 40 years of US experience. After generating the JSON format files for all datasets, we used OpenCV-Python (version 4.8.1.78) to calculate the bounding rectangle of the annotated ROI.

For OvaMTA-Seg, to locate the ovaries or nodules in the overall image, we directly used the complete US image, binary mask of the ROI, and pathology labels for supervised learning. For OvaMTA-Diagnosis, to focus more accurately on ROIs, we expanded their bounding rectangles by 25 pixels in each direction and cropped the US images to obtain ovarian mass patches. Pathology labels were processed using three-dimensional one-hot encoding, and the masks were processed as binary images that were proportional to the image.

The images used in this study had a range of sizes because different US devices were used to gather the images. Therefore, we uniformly adjusted all input images to 352 × 352 pixels by resizing them during training and validation. To reduce overfitting and increase the number and diversity of training sets, we applied the following transformations to images and masks for data augmentation: horizontal flip with a probability of 0.5, vertical flip with a probability of 0.5, and random rotation in the clockwise direction within 90°.[27,28]

Model architecture and training

This pipeline is composed of OvaMTA-Seg and OvaMTA-Diagnosis as OvaMTA, where OvaMTA-Seg roughly segments the ROIs in the US images and outputs the probability of the presence of ovarian masses. When there are no ovarian masses, our system directly outputs the segmentation result of OvaMTA-Seg and the probability of normal ovaries being present. When there are ovarian masses, our system crops and normalises all connected domains of the ROI separately and inputs the corresponding image patches into OvaMTA-Diagnosis for precise segmentation of the mass area and classification as benign or malignant. Finally, for the image, the detection box generated by the segmentation of the ovarian mass area and the corresponding malignant probability of the ovarian mass are outputted, as shown in Fig. 2.

The two models of the proposed system were trained separately. Considering the functionality of each model, the training set of OvaMTA-Seg included

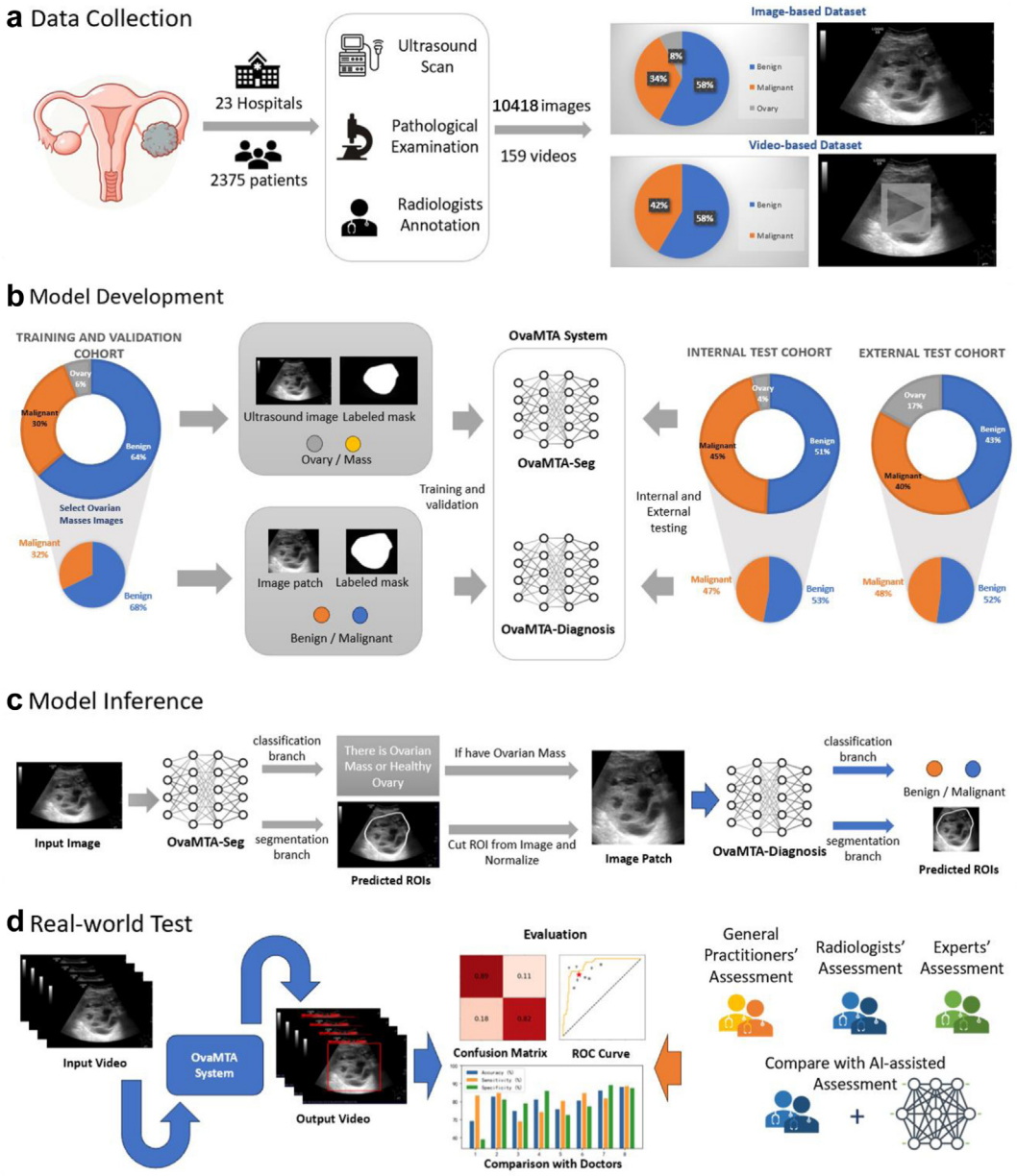

Fig. 2: Overview of the study design. (a). Collect ultrasound data and pathological results from 23 hospitals, and label the edges of ovarian masses. (b). Compose an image dataset and a video dataset for model development. (c). The OvaMTA system we propose consists of two neural networks: the OvaMTA-Seg network for automatic tumor detection and the OvaMTA-Diagnosis network for predicting malignancy. (d). Finally, we evaluate model performance using ROC, confusion matrix, accuracy, sensitivity, specificity, and Kappa coefficient, and compare with doctors in video testing. ROC, receiver operating characteristic curve. OvaMTA, ovarian multi-task attention network. ROI, region of interests. AI, artificial intelligence.

US images of all training cases of ovaries and ovarian masses, whereas the training set of OvaMTA-Diagnosis only included US images of all ovarian masses of all training cases (Fig. 2). Both models were developed based on MTANet.[27] The is a one stage multi-task attention network which includes a Pyramid Visual Transformer (PVT) for classification and a UNet-like multi-attention decoder for segmentation, which we previously developed. It outperformed state-of-the-art models in the CVC ClinicDB dataset for polyp segmentation, the ISIC-2018 dataset for skin lesion segmentation, and the US dataset for liver tumour segmentation and classification,[27] so we used it as the backbone of OvaMTA-Seg and OvaMTA-Diagnosis.

Based on the output probability values of the OvaMTA for the training and validation sets, we used a two-step decision method and selected two thresholds. When the probability of having ovarian masses in the image was less than 0.9, the image included only healthy ovaries and outputted the segmentation results of the ovary. In contrast, when the probability was higher than 0.9, the image was predicted to contain ovarian masses, and the connected domains of the segmented area generated by OvaMTA-Seg were cropped by bounding rectangles and entered into OvaMTA-Diagnosis for the precise segmentation of masses and diagnosis of benign and malignant masses. When the malignant probability of the OvaMTA-Diagnosis output is higher than 0.37, the model diagnoses the corresponding patch as malignant; otherwise, the patch is diagnosed as benign. It should be emphasised that this set of thresholds was selected empirically. It is possible that the optimal threshold for the same diagnostic model may not be the same in different test populations.[29] The optimal threshold mentioned here refers to the maximum value between the true positive rate and the false positive rate at this threshold. The threshold selected in this manner can, to some extent, balance sensitivity and specificity.[30]

The OvaMTA-Diagnosis and OvaMTA-Seg models were implemented using PyTorch on an NVIDIA GeForce RTX 3080 GPU. A batch size of 10 and AdamW optimiser with a learning rate of 10-5 were used for all models. We trained each model for 200 epochs using a 0.1 decay learning rate scheduler with 30 epochs. A validation set containing 10% of the images in the training and validation datasets was randomly selected to optimise the network and set up the hyperparameters. Eventually, OvaMTA outputs ROIs' bounding boxes and their probabilities of healthy ovary, benign, or malignant tumours, with the sum of these three probability values equal to 1, and all probability values distributed within the range 0–1. The average Dice coefficient was used as a metric to evaluate the performance of the model in segmentation (Supplementary Table S2). We used various metrics to assess the model's performance on classification tasks, including sensitivity (SEN), specificity (SPE), accuracy (ACC), Cohen's kappa score, and area under the receiver operating characteristic curve (AUC).

The implementation details of the deep learning framework are available at: https://github.com/SLYXDWL/OvaMTA.

Generalizing the model to real-world ultrasound diagnostic scenarios

For image-based predictions at the individual level, the average values of the predicted probabilities of multiple images for each individual were combined into a single score. For video-based prediction, the video is normalised frame-by-frame and then inputted into OvaMTA, which outputs the diagnostic result of the current frame in real time, and finally outputs the benign and malignant diagnosis of the entire video. To take contextual information into consideration, we average feature maps of previous six frames as feature map of present frame, and output binarised segmentation regions with a threshold of 0.5. The current output segmentation area is expanded by 25 pixels to take a patch, and then output the probability of being healthy, benign, and malignant for the corresponding frame. After outputting the probability for each frame, the frame-level malignancy probabilities of the video were averaged to obtain the probability of being benign or malignant for the entire video.

Comparison with radiologists and performance enhancement by AI

Six radiologists and two general practitioners were recruited to assess ovarian masses, including two junior radiologists (≥3 years clinical-experience), two intermediate radiologists (≥5 years clinical-experience), two senior radiologists (≥10 years clinical-experience), and two non-radiological practitioners (≥3 years clinical-experience). The initials and detailed experience of each doctor are included in Supplementary Table S3. The eight doctors independently evaluated all anonymised and randomised videos at the patient level in external dataset B and assigned each mass a benign or malignant label. After 2 weeks, the OvaMTA results were provided to radiologists for re-evaluation. The edges of the tumour segmented using the OvaMTA system and the bounding box with the probability of malignancy were displayed side by side with the original US video to the doctors (Fig. 5). The video-level diagnostic results of OvaMTA will also be presented to doctors at the end of the videos. All readers were blinded to the pathological diagnosis.

Statistics

In the development of our model, encompassing our training, validation, and internal testing datasets, we employed a retrospective analysis approach to integrate pathological information and follow-up data. The participants were allocated to the training, validation, and internal testing groups in a ratio of 7:1:2, yielding 1514 individuals in the combined training and validation group and 363 individuals in the internal testing group. When assessing the model, which comprises two separate testing cohorts, we adhered to the previously mentioned inclusion and exclusion criteria and assembled 339 participants for external testing set A and 159 participants for external testing set B. The 95% confidence interval (CI) of sensitivity, specificity, accuracy, Cohen's kappa score and AUC of the OvaMTA system and doctors were computed using bootstrapping. The sensitivity, specificity, and accuracy of the AI system were compared with those of doctors and AI-assisted

doctors using McNemar's test for paired categorical data. Statistical significance was set at $p < 0.05$. All analyses were performed using Python (version 3.10.12), and the packages used in this study included Scikit-learn (version 1.4.1. post1) and SciPy (version 1.11.2).

Role of funding source

The funders had no role in the design and conduct of the study; the collection, management, analysis, and interpretation of the data; the preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. W. Dai and Y. Wu accessed the data and had the final responsibility for the decision to submit it for publication.

Results

Demography of the participants

A total of 2375 patients (10418 images, 159 videos) were included in this study (Fig. 1). The data collected from the 21 hospitals were split as follows for model development: training and validation set (6938 images, 1514 patients) and internal test set (1584 images, 363 patients). The data collected from another two hospitals patients (1896 images, 339 patients) were included as external test set A. Demographic and clinical characteristics of the patients are shown in Table 1. All patients provided age information. In the training and validation datasets, the age of the patients ranged 12–86 years (median = 41), and in the internal test datasets, the age of the patients ranged 15–81 years (median = 45). In the external test image and video datasets, the age of the patients ranged 11–75 years old (median = 39) and 17–81 (median = 43 years), respectively.

Performance of AI system in ovary and ovarian masses segmentation

We tested the Dice coefficient of the segmentation model for both internal and external image-based test sets. The OvaMTA-Seg model achieved an average dice coefficient of 0.887 ± 0.134 in the internal test set and 0.819 ± 0.201 in the external test set A, as shown in Supplementary Table S2. The Dice coefficient of benign nodules in the internal test set is 0.912 ± 0.103, and the dice coefficient in the external test set is 0.843 ± 0.192. The dice coefficient of malignant tumours in the internal test set is 0.878 ± 0.131, and in the external test set it reaches 0.851 ± 0.141.

Fig. 5 shows an example of the segmentation results of OvaMTA-Seg for US images. AI frameworks can be used to perform precise US image segmentation. In video testing, the model can complete the entire detection and tracking process from the appearance to disappearance of the tumour. Therefore, the segmentation system alone can be used by radiologists as a visualisation tool to highlight lesion areas.

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. There were 2375 patients (10418 images, 159 videos) in the study. Mean data are ±SD with the range in parentheses. Categorical variables are shown as numbers; data in parentheses are percentages. SD, standard deviation. [a]Mixed histology of ovarian mass indicated the mass contains multiple pathological subtypes simultaneously. [b]Other histology encompassed mesenchymal tumor, Brenner tumor, etc. Samples include images and videos.

Table 1: **Baseline characteristics of patients.**

Performance of AI system in ovarian masses detection

OvaMTA followed a two-stage decision-making process; therefore, we conducted diagnostic tests for the presence of masses and benign/malignant lesions in both the internal and external image-based test sets. In terms of tests on the presence of masses, the OvaMTA-Seg model achieved an AUC of 0.970 (95% CI: 0.970, 0.973) in the internal test set and 0.877 (95% CI: 0.877, 0.882) in the external test set A. In terms of masses detection, considering Intersection over Union (IoU) > 0.5 as successful detection, the internal test set achieved a detection rate of 94.2, whereas the external test set achieved a detection rate of 88.1 based on images. In the external test set B, the model achieved a 100% detection rate in video-based testing, in which all the masses in the video were detected.

For the detection of malignant masses, our model achieved an AUC of 0.941 (95% CI: 0.940, 0.942) in the internal test set and 0.941 (95% CI: 0.938, 0.941) in the external test set A (Fig. 3). Notably, in the video case of external test set B, our model achieved an AUC of 0.911 (95% CI: 0.909, 0.913) (Fig. 4.) and an accuracy of 86.2 (137/159; 95% CI: 80.5, 91.2) in the benign and malignant classification task.

Comparison of diagnostic performance between AI system and doctors

As shown in Table 2, in the external test set B, the senior radiologists' assessment has an accuracy of 88.1 (95% CI: 84.3, 91.5), a sensitivity of 88.6 (95% CI: 82.9, 93.9), and a specificity of 87.6 (95% CI: 82.6, 92.3). The junior radiologists' assessment has an accuracy of 74.8 (95% CI: 70.1, 79.6), a sensitivity of 68.9 (95% CI: 61.0, 76.9), and a specificity of 79.0 (95% CI: 73.0, 84.7). The intermediate radiologists assessment has an accuracy of 75.8 (95% CI: 71.1, 80.5), a sensitivity of 80.3 (95% CI: 73.1, 86.9), and a specificity of 72.6 (95% CI: 66.3, 79.2). The AI system give an assessment comparable to senior radiologists, with an accuracy of 86.2 (137/159; 95% CI: 80.5, 91.2; $p = 0.50$, a sensitivity of 81.8 (54/66; 95% CI: 72.3, 91.0; $p = 0.16$), and a specificity of 89.2 (83/93; 95% CI: 82.8, 95.2; $p = 0.68$). The AI system outperformed intermediate radiologists ($p = 0.00026$), junior radiologists ($p < 0.0001$), and general practitioners ($p < 0.0001$). The sensitivity of the AI system was better than that of the junior radiologists ($p = 0.0046$). The specificity of the AI system was higher than that of intermediate radiologists ($p < 0.0001$), junior radiologists ($p = 0.0019$) and general practitioners ($p < 0.0001$). Overall, the diagnostic ability of the AI system was superior to that of mid- and low-seniority radiologists and comparable to that of high-seniority radiologists.

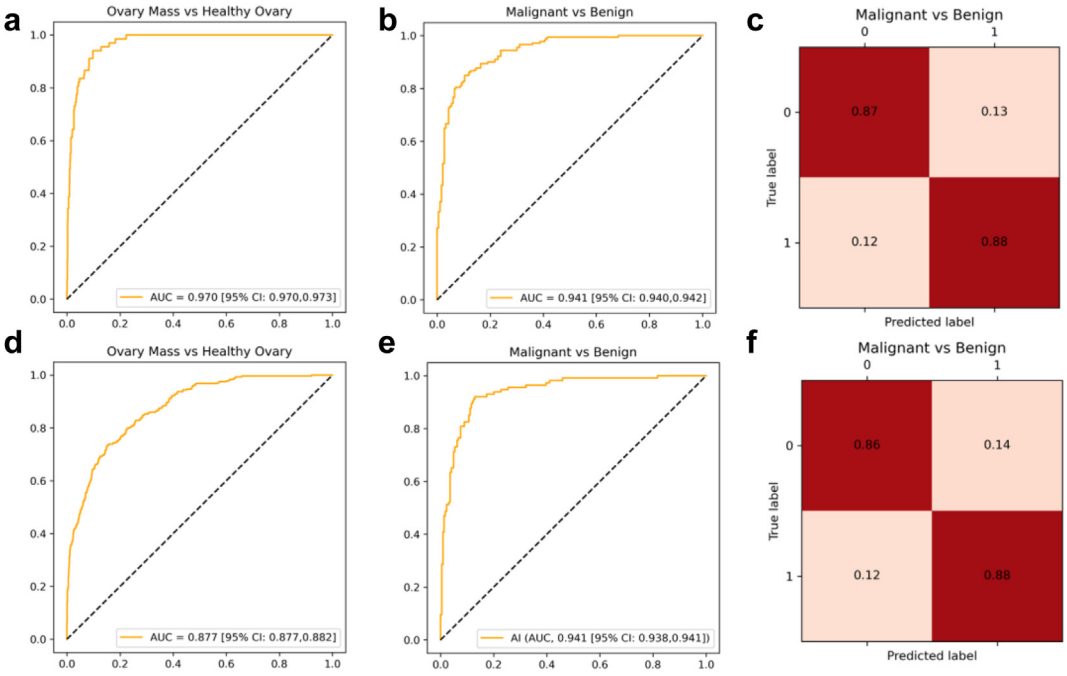

Fig. 3: Diagnostic performances of OvaMTA in image-based testing. ROC and confusion matrices of internal test (a, b, c) and external test A (d, e, f) datasets shows the diagnostic performance of AI system in detecting ovarian masses and discriminating benign and malignant ovarian tumors based on images. ROC, receiver operating characteristic curve. OvaMTA, ovarian multi-task attention network. AI, artificial intelligence.

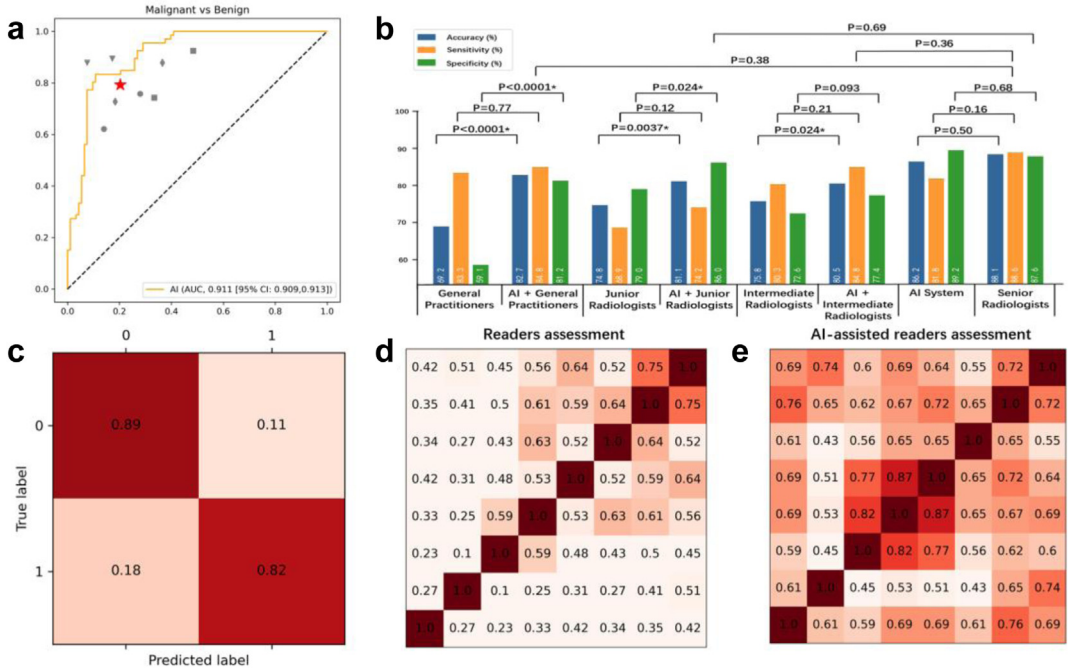

Fig. 4: Comparison of diagnostic performance between AI system, radiologists, and general practitioners. (a). ROC of external test videos shows the diagnostic performance of AI system. And comparing with radiologists of different experience levels. Big red star represents the average value of radiologists. The grey circle represents the performance of junior radiologists, the grey diamond represents the performance of intermediate radiologists, the gray triangle represents the performance of expert, and the grey square represents the performance of general practitioners. (b). Bar plot shows performances of eight doctors with indicated paired McNemar test *p*-values for group comparison; (c). Normalized confusion matrix of AI system's performance in discriminating benign and malignant ovarian tumors in external test videos. (d). Kappa matrix of eight doctors' assessment in external test videos. (e). Kappa matrix of eight doctors' assessment assisted with AI in external test videos. Each value in the matrices represents the inter-observer agreement between two doctors. On the axes of kappa matrices, general practitioners 1–2 and radiologists 1–6 are arranged in the order from bottom to top and from left to right in (e), (d). ROC, receiver operating characteristic curve. AI, artificial intelligence.

Auxiliary diagnosis function of AI system on ultrasound videos

With the assistance of OvaMTA, the accuracy of the general practitioners significantly improved from 69.2 to 82.7, $p < 0.0001$. The accuracy of junior radiologists was improved from 74.8 to 81.1, $p = 0.0037$, and the accuracy of intermediate radiologists was improved from 75.8 to 80.5, $p = 0.024$. And there was a significant improvement in junior and intermediate radiologists who were assisted by AI compared to those who were not assisted by AI (ACC: 80.8 vs. 75.3, $p = 0.00015$; SEN: 79.5 vs. 74.6, $p = 0.029$; SPE: 81.7 vs. 75.8, $p = 0.0032$). AI system significantly enhances the specificity of junior doctors, including general practitioners 59.1–81.2, $p < 0.0001$, junior radiologists 79.0–86.0, $p = 0.024$. The specificity of these junior radiologists assisted by AI has reached a level comparable to that of senior radiologists, 86.0 vs. 87.6, $p = 0.69$. And the sensitivity of non-radiological doctors assisted by AI has reached a level comparable to that of senior radiologists, 84.8–88.6, $p = 0.38$. With the assistance of AI, general practitioners reached the average level of radiologists (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 Kappa matrices calculated and visualised in Fig. 4 (d) and (e) show that, with the assistance of AI, the diagnosis between different doctors has become more consistent. Therefore, the AI system improved the diagnostic level of doctors as well as their interobserver agreements.

Visualization and interpretability of AI system

As shown in Fig. 5 and the attached videos, OvaMTA can not only accurately locate the tumour, but also discover the edge ovarian tissue of the tumour and perform reasonable analysis on each frame of the video. Our model can assist doctors in diagnosing benign and malignant ovarian masses. Videos S1 and S2 show that the AI system stably and continuously outputs frame-by-frame segmentation and diagnostic results.

Fig. 5 (a) shows a keyframe of a greyscale US video (Video S1) of a 64-year-old woman with solid component(s) within a multilocular cyst. The AI system can

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. Data in parentheses are numerators/denominators; data in brackets are 95% CIs. Detailed results are presented in Supplementary Tables S3; *$p < 0.05$, Significant difference with AI system. AI, artificial intelligence. OvaMTA, ovarian multi-task attention network. CI, confidence interval. [a]The differences between readers and AI system were compared, and p values were calculated.

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

detect and map the edges of a mass and perform malignant classifications and probabilities in real-time. The final diagnostic result provided by the AI system was consistent with the postoperative pathological results (endometrioid carcinoma). The AI system provides interpretable heat maps in real-time, allowing doctors to pay more attention to irregular compartments and solid components. Fig. 5 (b) shows a keyframe of a greyscale US video (Video S2) of a 52-year-old woman with a unilocular anechoic cyst. The AI system can detect and map the edges of the mass and make benign classifications and probabilities in real-time. The final diagnosis provided by the AI system was consistent with the postoperative pathological results (serous cystadenofibroma). The AI system provides interpretable heat maps in real-time, allowing doctors to pay more attention to the anechoic cystic part.

Discussion

Accurate identification and diagnosis of ovarian cancer are of significance for improving the survival rate and prognosis of patients with ovarian cancer. Recent developments in deep learning methods have demonstrated their feasibility in ovarian cancer diagnosis using ultrasound (US) images. However, there is still no AI pipeline to diagnose ovarian cancer that meets the actual clinical workflow. Therefore, we designed an automated deep learning pipeline OvaMTA, for ovarian mass detection and prediction of pathological type based on US screening, and tested it on images as well as videos. The model performed well for ovarian mass localisation and ovarian localisation in the ultrasonic images (mDice: 0.819). At the same time, it had excellent diagnostic ability for benign and malignant ovarian masses in ultrasonic scanning videos (AUC: 0.911). The OvaMTA system based on ultrasound imaging is a simple and practical auxiliary tool for screening for ovarian cancer, with a diagnostic performance comparable to that of senior radiologists. This provides a potential tool for detecting and diagnosing ovarian cancer.

OvaMTA is an integrated system for the detection, segmentation, and diagnosis of ovaries and ovarian masses. In the scanning process for ovarian cancer, it is essential to detect regions of interest in US imaging; radiologists can only obtain complete size measurements and observations to evaluate the physiological status of the ovary or ovarian mass. Previous studies have focused only on ovarian masses,[10,12,14–18,31,32] with limited training data or a lack of external testing. The corresponding model was also more suitable for use after checking and retaining all images and clinical information, making it difficult for it to play a role in busy clinical scenarios. For ease of use, it is necessary to develop an automated segmentation module suitable for ovarian US imaging. This not only improves AI accuracy but also improves the interpretability of AI to a certain extent.

OvaMTA achieved an advanced performance. In this study, we used a multicentre dataset, healthy ovarian labels, and an updated network structure to improve workflow. The backbone model MTANet outperforms ResNet and DensNet,[33,34] which are commonly used in the diagnosis of ovarian cancer in medical imaging,[16,18] because it can organically combine segmentation and classification tasks, enabling the two tasks to promote each other and achieve higher accuracy.[27] Diversified training data can also promote models to adapt to different populations on a larger scale. Our system achieved the best mIoU score than existing studies,[35,36] and demonstrated higher accuracy than previous studies reported in a meta-analysis (SEN: 0.90 vs. 0.88, SPE: 0.86 vs. 0.84).[37]

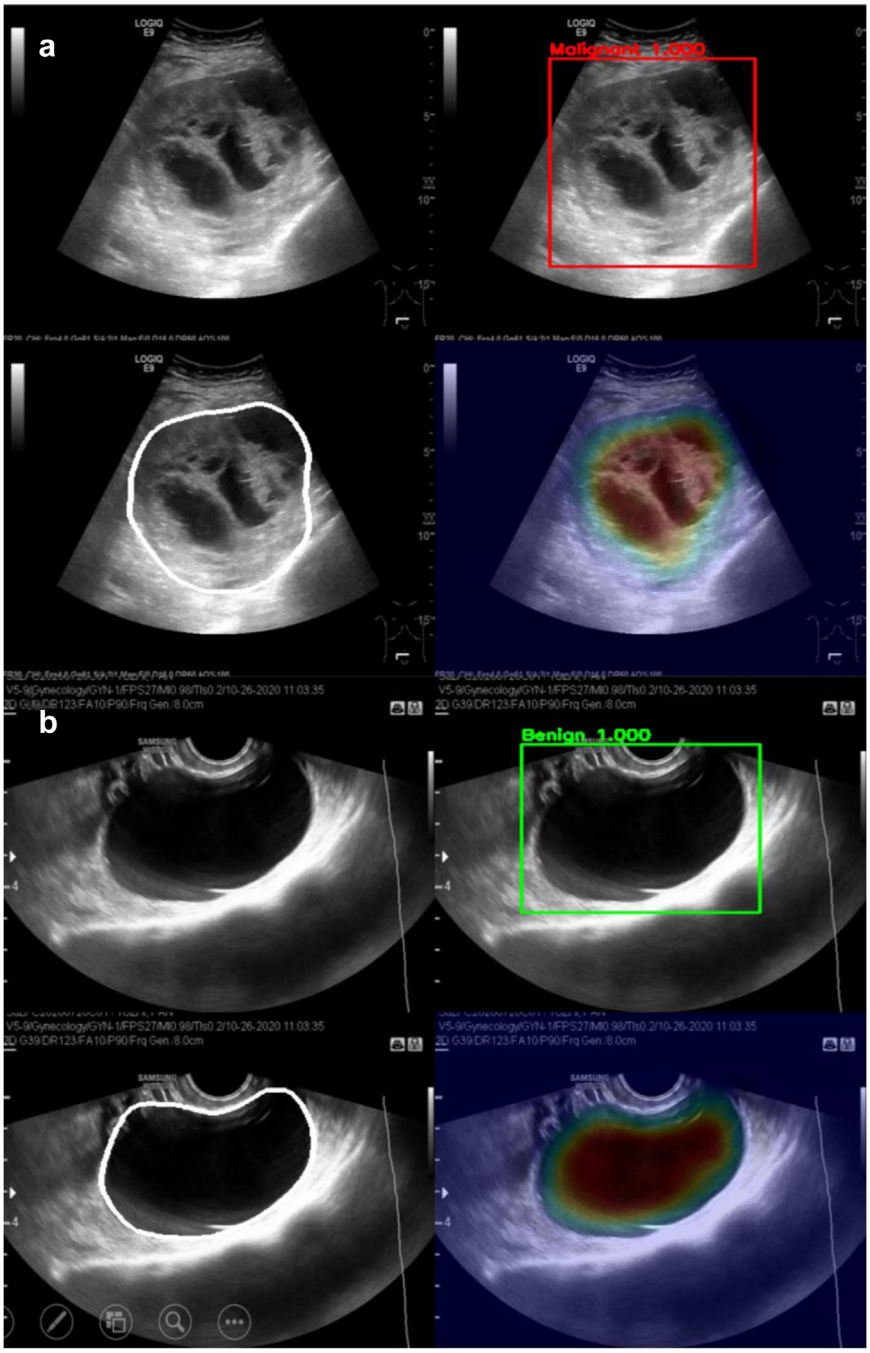

Fig. 5: Two types of ovarian masses detected, segmented, diagnosed, and interpreted by AI system. (a). Keyframe of greyscale ultrasound videos in a 64-year-old woman with solid component(s) within a multilocular cyst. The AI system can detect and map the edges of the mass, and make malignant classification and probability in real time. The final diagnosis result provided by AI system was consistent with the pathological result after surgery (endometrioid carcinoma). The AI system gives interpretable heat maps in real time, allowing doctors to pay more attention to the irregular compartments and solid components. (b). Keyframe of greyscale ultrasound videos in a 52-year-old woman with a unilocular anechoic cyst. The AI system can detect and map the edges of the mass, and make benign classification and probability in real time. The final diagnosis result provided by AI system was consistent with the pathological result after surgery (serous cystadenofibroma). The AI system gives interpretable heat maps in real time, allowing doctors to pay more attention to the anechoic cystic part. AI, artificial intelligence.

OvaMTA has demonstrated strong generalisation and robustness and is highly clinically applicable. In this study, we used multicenter data for external testing and conducted video-level testing, which provided a more comprehensive validation of the model's generalisation and stability compared with other studies. To achieve high diagnostic efficiency, our model also implements a fully automated process, which makes it easy to embed in the workflow of doctors, no longer just a tool for verifying and improving diagnostic efficiency after diagnosis. Ovarian masses are located deep within the pelvic cavity, making their detection and diagnosis of ovaries and masses challenging. AI systems can assist doctors in detecting the ovaries and masses. After segmentation, the tumour was further classified and diagnosed, and a probability value was provided as a more intuitive diagnostic tool for doctors. To enhance interpretability, we generated a real-time attention heat map based on the feature map of the current frame, which enabled doctors to better understand the focus area of the model (Videos S1 and S2). The AI system not only significantly improves the accuracy of junior radiologists from 74.8 to 81.1, in the diagnosis of benign and malignant ovarian masses during screening but also has the potential to assist radiologists in conducting US examinations. In addition, after a short training for triage and primary diagnosis in outpatient clinics, the AI system can also assist non-ultrasound specialised general practitioners achieve the average level of radiologists (ACC: 82.7 vs. 81.8, $p = 0.80$; SEN: 84.8 vs. 82.6, $p = 0.72$; SPE: 81.2 vs. 81.2, $p > 0.99$). Therefore, AI systems can be used in more application scenarios such as screening in grassroots hospitals.

Our study had some limitations. First, by observing external testing in multiple regions (Supplementary Table S4), we identified the possibility of a threshold shift in AI. Therefore, the thresholds mentioned in this work are only a reference with good performance, and using the domain generalisation technique to optimise the AI model is meaningful.[38] Second, because of the inclusion of only tumour scanning videos from patients with ovarian masses, we did not test the effectiveness of the model in segmenting normal ovaries in the videos and only tested it on static images. It is important to identify "normal ovary" during ovarian cancer screening, especially in patients after menopause. Our work does not include imaging of other pelvic masses, nor has it included imaging of multiple targets (multiple masses, or masses coexisting with normal ovaries). Adding negative and normal samples such as health adults to training is crucial for the functionality of the model, especially the models planned to be used in screening scenarios in the future.[39] Therefore, targeted data collection and model training should be conducted in the future. Third, we observed that imprecise segmentation may lead to a decrease in diagnostic performance (Supplementary Table S5), and there is room for improvement in the collaborative mode between humans and AI.[40,41] Fourth, the current pipeline lakes the keyframe determination model, which leads to bias in video-level judgment. The final video output of the model may be affected by non-keyframes, resulting in incorrect overall predictions. Owing to the lack of dense, frame-by-frame benign and malignant labels, such labelling is not in line with the habits of doctors, which may lead to incomplete performance evaluation. Therefore, it is important to develop a decision algorithm that adapts to human scanning habits. Finally, colour Doppler imaging and clinical staging diagnoses were not conducted in this study because doctors have different definitions and conceptual understandings of colour Doppler, which can easily result in bias. We plan to further explore clinical staging and multimodal imaging to provide more precise US diagnostic tools for ovarian masses.

In conclusion, OvaMTA based on integrating ultrasonic images can detect and classify healthy ovaries and ovarian masses, and further diagnose benign and malignant ovarian masses, and its diagnostic performance is comparable to that of expert subjective assessment. AI systems can significantly improve the diagnostic abilities of junior and intermediate radiologists. General practitioners assisted by an AI system can achieve an average performance of radiologists. OvaMTA has the potential to be an effective and widely applicable tool for assisting doctors in the screening of ovarian masses and diagnosis for ovarian cancer.

Contributors

Guarantors of integrity of entire study, W.L.D., Y.N.W., D.X.K., L.T. S.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, W.L.D., Y.N.W., clinical studies, W.L.D., Y.N.W., J.Z., L.P.G., Z.W.G., S.Z., M.N.Z., S.D., S.C.X., L.W., L.T.S.; experimental studies, W.L.D., Y.N.W., Y.T.L., J.Z., L.P.G., Z.W.G., S.Z., M.N.Z., S.D., S.C.X., L.W., L.T.S.; statistical analysis, W.L.D., Y.N.W., Y.T.L.; and manuscript editing, D.X.K., L.T.S.

W.L.D. and Y.N.W. have verified the underlying data. All authors read and approved the final version of the manuscript, and ensure it is the case.

Data sharing statement

As the study involved human participants, the data cannot be made freely available in the manuscript or in a public repository because of ethical restrictions. However, the data are available from the Zhejiang Provincial People's Hospital to researchers who meet the criteria for access to confidential data. Interested researchers can send data access requests to the corresponding author (L. T. S.). The code of the deep learning framework is available at: https://github.com/SLYXDWL/OvaMTA.

Declaration of interests

All authors declare no competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 12090020, 82071929 and 12090025) and the R&D project of Pazhou Lab (Huangpu) (Grant Nos.2023K0605).

We appreciate Q.X. J.'s contribution as a general practitioner in the AI assisted interpretation experiment of this work. We appreciate Y.P.B.'s contribution in data annotation and image quality control. Their contributions have been fundamental to the development and advancement of this research.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102923.

References

1 Bray F, Laversanne M, Sung H, 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):229–263.
2 Zheng R, Zhang S, Zeng H, 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, 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):448–458.
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):922–929.
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):285–291.
7 Timmerman D, Testa AC, Bourne T, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol*. 2008;31(6):681–690.
8 Van Calster B, Van Hoorde K, Valentin 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, 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):168–185.
10 Kongara SR, Prakasha S, Brindha A, et al. Performance evaluation of optimized convolutional neural network mechanism in the detection and classification of ovarian cancer. *Multimed Tool Appl*. 2024. https://doi.org/10.1007/s11042-024-18115-0.
11 Xiang H, Xiao Y, Li F, 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, 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, et al. Prediction of benign and malignant ovarian tumors using Resnet 34 on ultrasound images. *J Obstet Gynaecol Res*. 2023;49(12):2910–2917.
14 Jung Y, Kim T, Han MR, 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, 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, et al. Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. *Radiology*. 2022;304(1):106–113.
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;11:110681–110691.
20 C K, Shivaram JM. Segmentation of ovarian cyst using improved U-NET and hybrid deep learning model. *Multimed Tool Appl*. 2023;83(14):42645–42679.
21 Li J, Chen Y, Zhang M, 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, et al. Radiomics analysis of ultrasound images to discriminate between benign and malignant adnexal masses with solid ultrasound morphology. *Ultrasound Obstet Gynecol*. 2024. https://doi.org/10.1002/uog.27680.
23 Fanizzi A, Arezzo F, Cormio G, 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, 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, et al. CR-unet: a composite network for ovary and follicle segmentation in ultrasound images. *IEEE J Biomed Health Inform*. 2020;24(4):974–983.
26 Herrington CS. *WHO classification of tumours female genital tumours*. 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):674–685.
28 Kim T,LD, 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 Computer Science*. 2020;1(2).
32 Barcroft JF, Linton-Reid K, Landolfo 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. 30th ieee conference on computer vision and pattern recognition (cvpr 2017)*. 2017:2261–2269.
34 He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. *Proc Cvpr Ieee*. 2016:770–778.
35 Qi Zhaoa SL, Baib W, Caia L, et al. MMOTU: a multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *arXiv*. 2023. https://doi.org/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)*. 2023:1.
37 Xu HL, Gong TT, Liu FH, 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, et al. Generalizing to unseen domains: a survey on domain generalization. *IEEE Trans Knowl Data Eng*. 2022;35(8):8052–8072.
39 He Y, Guo Y, Lyu J, 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, et al. Generalizability and diagnostic performance of AI models for thyroid US. *Radiology*. 2023;307(5):e221157.
41 He B, Kwan AC, Cho JH, et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature*. 2023;616(7957):520–524.