Journal of Ovarian Research

RESEARCH

Open Access

Machine learning models in evaluating the malignancy risk of ovarian tumors: a comparative study

Xin He[1,2], Xiang-Hui Bai[3], Hui Chen[1*] and Wei-Wei Feng[1*]

Abstract

Objectives The study aimed to compare the diagnostic efficacy of the machine learning models with expert subjective assessment (SA) in assessing the malignancy risk of ovarian tumors using transvaginal ultrasound (TVUS).

Methods The retrospective single-center diagnostic study included 1555 consecutive patients from January 2019 to May 2021. Using this dataset, Residual Network(ResNet), Densely Connected Convolutional Network(DenseNet), Vision Transformer(ViT), and Swin Transformer models were established and evaluated separately or combined with Cancer antigen 125 (CA 125). The diagnostic performance was then compared with SA.

Results Of the 1555 patients, 76.9% were benign, while 23.1% were malignant (including borderline). When differentiating the malignant from ovarian tumors, the SA had an AUC of 0.97 (95% CI, 0.93–0.99), sensitivity of 87.2%, and specificity of 98.4%. Except for Vision Transformer, other machine learning models had diagnostic performance comparable to that of the expert. The DenseNet model had an AUC of 0.91 (95% CI, 0.86–0.95), sensitivity of 84.6%, and specificity of 95.1%. The ResNet50 model had an AUC of 0.91 (0.85–0.95). The Swin Transformer model had an AUC of 0.92 (0.87–0.96), sensitivity of 87.2%, and specificity of 94.3%. There was a statistically significant difference between the Vision Transformer and SA, and between the Vision Transformer and Swin Transformer models (AUC: 0.87 vs. 0.97, $P=0.01$; AUC: 0.87 vs. 0.92, $P=0.04$). Adding CA125 did not improve the diagnostic performance of the models in distinguishing benign and malignant ovarian tumors.

Conclusion The deep learning model of TVUS can be used in ovarian cancer evaluation, and its diagnostic performance is comparable to that of expert assessment.

Keywords Ovarian cancer, Ultrasound, Machine learning, Diagnostic models

*Correspondence:
Hui Chen
ch11516@rjh.com.cn
Wei-Wei Feng
fww12066@rjh.com.cn
[1]Department of Obstetrics and Gynecology, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, P.R. China
[2]Department of Obstetrics and Gynaecology, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, Hong Kong, P.R. China
[3]Philips Health Technology (China) Co., Ltd. Shanghai Branch, 718 Lingshi Road, Shanghai 200072, P.R. China

Background

Ovarian Cancer (OC) is a major concern for women, with the highest mortality rate among gynecological cancers [1, 2]. Accurate classification of these groups prior to surgery is vital for determining appropriate treatment [3]. Precise laboratory test such as CA125, a protein biomarker, is commonly used in clinical practice to assess ovarian cancer [4]. Elevated levels of CA125 can indicate the presence of ovarian cancer, however, it is important to note that CA125 levels can be elevated in non-cancerous conditions as well, namely endometriosis or pelvic inflammatory disease and not all ovarian cancers produce high levels of CA125 [5]. As a consequence, ultrasound (US) is currently the preferred imaging modality for evaluating ovarian cancer due to its convenience, sensitivity, and affordability [6]. The great disadvantage of ultrasound is the strong operator dependence, an expert's subjective assessment is still the most reliable evaluation of adnexal pathology [7]. To resolve this issue, The diagnostic ultrasound approach has undergone significant advancements, transitioning from subjective experience-based evaluation to more structural evidence-based algorithms such as Simple Rules (SR), the ADNEX, LR1, LR2 risk models [8–11].

Machine learning, especially deep learning domain is a fascinating and powerful tool for computer vision. It becomes a promising and robust tool in ultrasound imaging classification, detection, and segmentation [12]. In a study by Christiansen et al., two innovative deep neural networks were constructed for diagnosing ovarian cancer [13]. Ovry-Dx1 achieved a sensitivity of 96.0% and a specificity comparable to clinical experts, while Ovry-Dx2 demonstrated a sensitivity of 97.1% and a specificity of 93%. Combined with expert evaluation, they significantly increased overall sensitivity (96.0%) and specificity (89.3%).Additionally, a collaborative study with 10 hospitals revealed that the machine learning model outperformed the average diagnostic level of radiologists matched the level of expert ultrasound image readers for ovarian tumors [14]. Furthermore, our previous research involving 422 patients found that the ResNet performed comparably to expert subjective assessments (SA) and the Ovarian-Adnexal Reporting and Data System [15].

In recent years, advancements in powerful hardware, new optimized techniques, software libraries, and large datasets has accelerated its growth and led to the emergence of new architectures such as the transformer. The Transformer, an attention mechanism-based model, has shown exceptional performance in various computer vision tasks, including tumor segmentation and classification [16]. In our study, we harnessed the potential of four cutting-edge deep learning pre-trained architectures, namely ResNet, DenseNet, Vision Transformer, and Swin Transformer, to differentiate the malignancy risk of ovarian tumors in ultrasound images and compare them to subjective assessment performed by an expert. Additionally, we explored the integration of CA125 for joint diagnosis purposes.

Methods

Patients

This single-center, retrospective, diagnostic accuracy study was conducted at the Department of Obstetrics and Gynecology at Ruijin Hospital in Shanghai, China, a tertiary referral oncology center. Between January 2019 and May 2021, 1,632 patients with an ultrasound diagnosis of an adnexal mass were consecutively enrolled. Inclusion criteria included the presence of at least one non-physiologic adnexal mass detected by transvaginal or transrectal ultrasonography, patient willingness to undergo surgery, less than 30 days between ultrasound and surgery, and no previous history of ovarian cancer. Exclusion criteria were histopathologic analysis–confirmed uterine sarcomas or non-gynecologic tumors, inconclusive histopathologic results, lack of medical records, or poor US image quality.

Data collection

Preoperative transvaginal ultrasonography was performed on all patients, with transabdominal ultrasound added if malignancy was suspected or if the mass was too large for transvaginal assessment alone. Ultrasound machines used were GE Voluson E10(GE Healthcare) and Philips IU22 and Philips A70 and EPIQ5(Philips Healthcare) with 5.0–9.0 MHz, and 3.0– 10.0 MHz transvaginal probes, respectively, and 1.0–5.0 MHz transabdominal probes. Clinical data including age, cancer antigen 125(CA125), pathologic results and ultrasonographic findings were recorded for each patient.

Subjective assessment

An experienced ultrasound expert (H.C.) with 11 years of clinical experience and 16 years of US experience assessed the sonographic tumor morphology according to the IOTA Group [10, 17].

In cases where multiple adnexal masses were present in a patient, the mass with the most complex ultrasound morphology was selected for risk estimation, if the masses had similar morphology, the largest tumor was chosen for inclusion in the study [10, 17]. The expert subjectively evaluated the malignancy of tumors, as following: 1, certainly benign; 2, probably benign; 3, uncertain but most likely benign; 4, uncertain but most likely malignant; 5, probably malignant; and 6, certainly malignant with criteria defined by Meys et al. [18].

Machine learning algorithm

In this paper, four deep leaning models were utilized for ovarian tumor risk stratification on ultrasound, which are:

- Residual Network(ResNet) [19]: ResNet introduces the concept of residual learning to address the degradation problem faced by very deep neural networks. The basic building block of ResNet is the residual block, which contains skip connections (shortcuts) that allow gradients to flow more directly during training. By using residual connections, ResNet can train very deep networks (e.g., hundreds of layers) without suffering from vanishing gradients or degradation in performance.
- Densely Connected Convolutional Network(DenseNet) [20]: DenseNet introduces dense connections between layers, where each layer receives direct input from all preceding layers and passes its own feature maps to all subsequent layers. Dense connections facilitate feature reuse and promote feature propagation throughout the network. By densely connecting layers, DenseNet encourages feature reuse, reduces the number of parameters, and enhances gradient flow, leading to improved performance and efficiency.
- Vision Transformer(ViT) [21]: ViT applies the transformer architecture, originally designed for sequence processing tasks like natural language processing (NLP), to image classification. ViT breaks down an image into fixed-size patches and flattens them into sequences, which are then fed into a transformer encoder. The transformer encoder processes these patches with self-attention mechanisms, capturing global dependencies and relationships between patches to make classification decisions. ViT has shown strong performance on image classification tasks, especially when pre-trained on large-scale datasets.
- Swin Transformer [22]: Swin Transformer is an extension of the transformer architecture specifically designed for vision tasks, aiming to handle both local and global dependencies efficiently. Unlike ViT, Swin Transformer adopts a hierarchical design with multiple stages, each containing a set of layers with local self-attention mechanisms. Swin Transformer employs shifted windows for self-attention computation, allowing it to capture both local and global information effectively. By leveraging hierarchical structures and shifted windows, Swin Transformer achieves strong performance on various vision tasks, including image classification, object detection, and segmentation.

For these four machine learning model development, we used Python 3.8 along with the PyTorch 2.1.2 deep learning library. Additionally, the models were pretrained on ImageNet-1 K dataset and finetuned with ovary ultrasound images. Three categories of US images were taken as input for the Deep learning(DL) algorithms, including gray scale US images depicting the plane with the maximum dimension and its orthogonal plane (two images per patient), color Doppler US images (one to three images per patient), and gray scale US images showing the maximum size of the solid component and its orthogonal plane (two images per patient if a solid component was present). In cases where there was no solid component, a blank image filled with zeros was used. The annotated images, where the region of the lesion and its solid component were manually segmented, were generated by the author (H.X.), using an open-source labeling tool (LabelMe).

To ensure unbiased results and model generalization, we followed a rigorous approach to divide the dataset into training, validation, and test sets. The dataset was stratified based on pathology results (benign vs. malignant) to ensure an even distribution of both benign and malignant cases across the subsets. We randomly split the data into training (80%), validation (10%), and test (10%) sets.

To further mitigate the risk of bias, we repeated the random splitting multiple times and evaluated the model performance on different random test sets. This approach ensured that the model performance was not reliant on any specific partition of the dataset.

Before input to the neural network, several preprocessing operations were applied to the original image which include:

- Crop: this operation is used to crop the region of ovary from the original ultrasound image,
- Resize: this operation resizes the cropped image to 256px x 256px;
- Remove caliper: this operation uses image processing method to remove measurement calipers burned on the image.

For model training, cross entropy loss were used with Adam optimizer. The learning rates were set from 1e-5 to 1e-4 and for different models, it took 50 to 100 epochs to train the models.

The image processing procedure is illustrated in Fig. 1. Three categories of US images were input to the network after preprocessing operations. DL models output the malignancy score for every input image and all these scores were averaged pooled to obtain the final prediction probabilities for each case. The final decision of benign or malignant was determined by comparing the

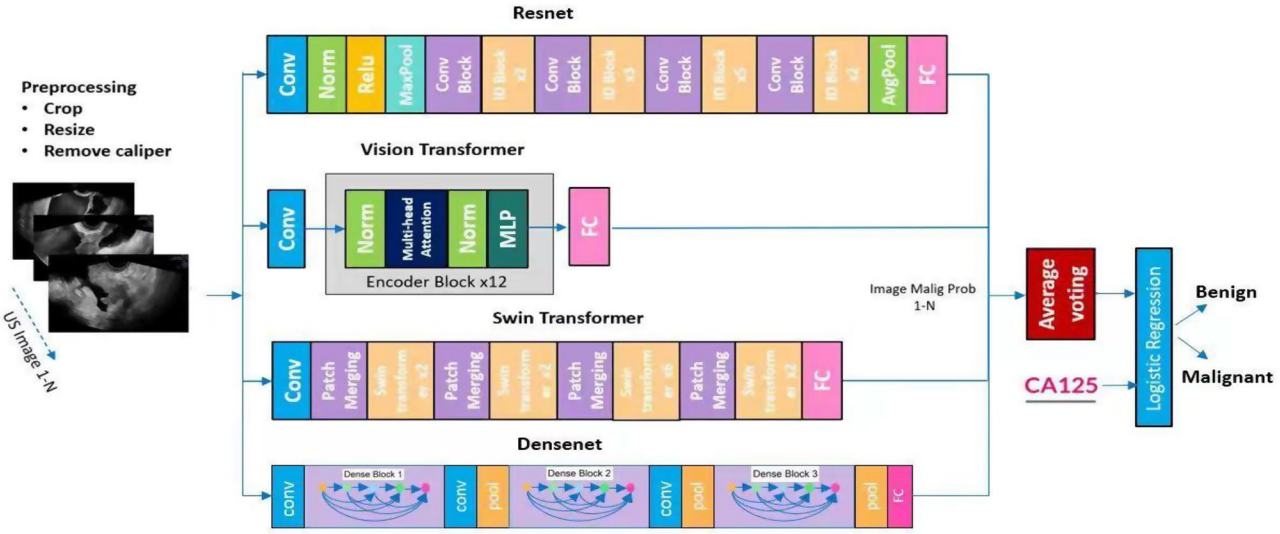

Fig. 1 Machine learning models flowcharts

output malignancy probability with a preselected cutoff threshold. This threshold aimed to achieve an optimal balance between sensitivity and specificity, maximizing the Youden Index value.

To better illustrate which part of the ultrasound image that most impact the classification result, this section utilizes Grad-CAM [23] to present heat maps depicting the regions of interest that the model concentrates on.

Reference standard

Histopathological diagnosis post-surgical removal was the reference standard. All patients underwent surgery, and final pathology results were obtained. Excised tissues were examined histologically according to the World Health Organization guidelines for tumor classification [24] and staged based on the International Federation of Gynecology and Obstetrics criteria [25]. In the final diagnosis, the masses were classified into two types: benign, and malignant, including BOT, Stage-I–IV OC and secondary metastatic cancer.

Statistical analysis

SPSS version 22.0 (IBM Corp) and MedCalc version 15.2.2 (MedCalc Software) were used for statistical analysis. Sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratiowere calculated. To compare the diagnostic performances among machine learning (ML) models and expert assessment, receiver operating characteristic curves(ROCs) were constructed and the areas under the receiver operating characteristic curves (AUCs) were calculated. Comparisons between AUCs were made by using the DeLong test. Cutoff values with optimal balance between sensitivity and specificity that maximize the Youden index in receiver operating characteristic curves were used to dichotomize the test set (i.e., the mass was classified as malignant when the scores extracted from ML models, and expert assessment were higher than the cutoff value). Tumor characteristics, patient features, and tumor marker levels were compared using appropriate statistical tests. All the statistical calculations were performed with 95% CIs and statistical significance was set at $P<0.05$. For the purposes of statistical analyses, borderline ovarian tumors were classified as malignant [26].

Ethical statement

This study was approved by the Ruijin Hospital, Shanghai Jiaotong University School of Medicine institutional ethics committee with exemption to obtain informed consent from individual patients (Grant No.2023-21). Written informed consent was waived due to the retrospective data collection. The study followed Good Clinical Practice (GCP) guidelines and the Netherlands Code of Conduct for Research Integrity.

Results

Patient characteristics

In this study, a total of 1,632 patients with adnexal tumors detected by ultrasound examination at the Department of Obstetrics and Gynecology, Ruijin Hospital affiliated to Shanghai Jiao Tong University School of Medicine between January 2019 and May 2021 were included. After applying exclusion criteria, 1,555 patients were analyzed, including 1,196 (76.9%) patients with benign tumors and 359 (23.1%) patients with malignant tumors. The flowchart of enrollment is shown in Fig. 2. Pathological results of the patients are summarized in Table 1, whereas demographic and clinical characteristics are presented in Table 2.

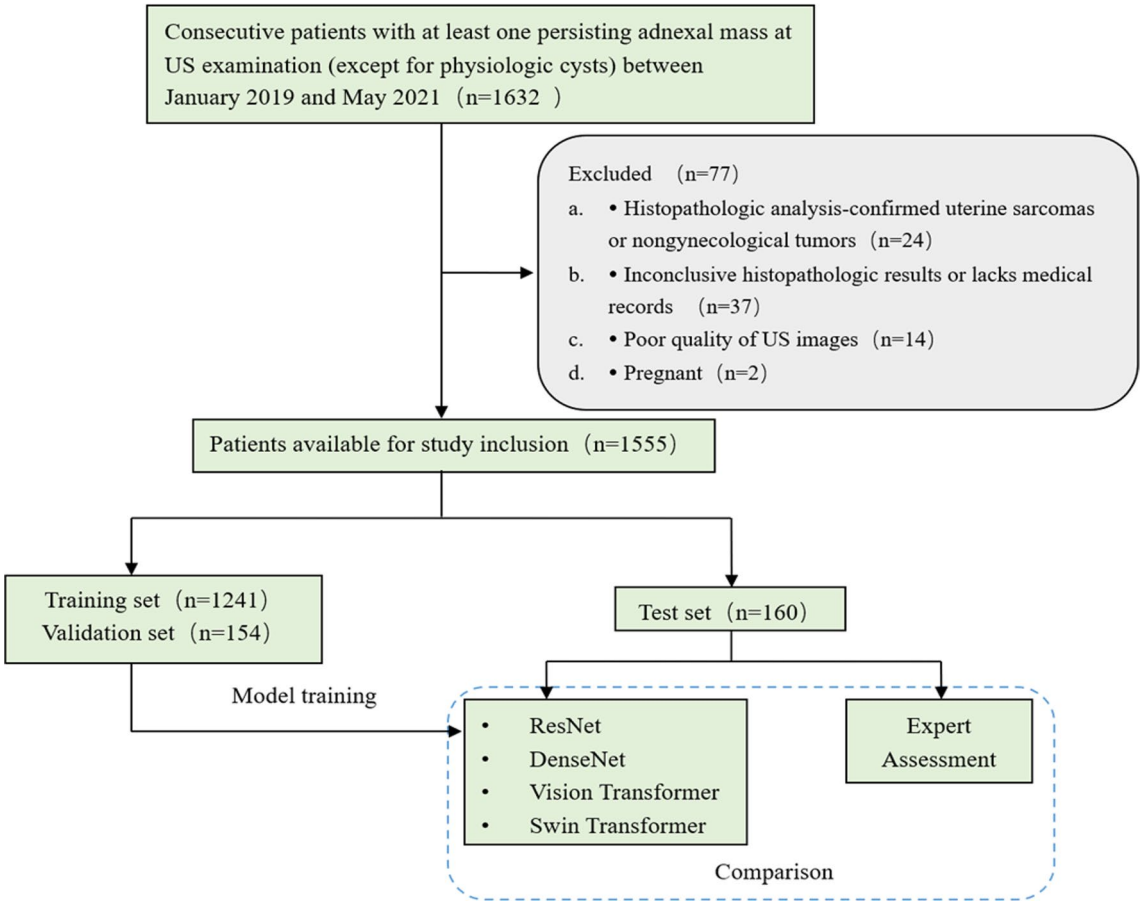

Fig. 2 Flowchart of enrollment in study cohort

The dataset was divided according to an 8:1:1 ratio, resulting in a training set (containing 956 benign and 285 malignant cases; totaling 7,493 images; 80%), a validation set (consisting of 119 benign and 35 malignant cases; comprising 799 images; 10%), and a test set (comprising 121 benign and 39 malignant cases; encompassing 818 images; 10%). Demographic and clinical characteristics between the training, validation, and test sets were consistent, as detailed in Table 3. There were no significant differences in age, CA125 levels, or other key clinical features, thus ensuring that the test set was representative of the patient population and reducing potential bias.

Significant differences were observed between benign and malignant tumors with respect to clinical and ultrasound characteristics. The mean age of patients with malignant tumors was higher than that of patients with benign tumors, with a median age at diagnosis of 54.0 and 41.0 years, respectively ($p<0.001$). Serum tumor markers showed significantly higher levels in patients with malignant tumors compared to those with benign tumors, as reflected by median values of CA125 (122.2 vs. 17.6, $p<0.001$). Ultrasound features also differed significantly between benign and malignant adnexal tumors. Malignant tumors had larger diameters for both mass and solid components (74 vs. 55 mm, $p<0.001$; 50 vs. 24 mm, $p<0.001$) and more abundant blood flow ($p<0.001$). There were also notable differences in tumor type between the two groups, with malignant tumors occurring more frequently in masses with solid component, while benign tumors were more likely to be simple cysts. Additionally, malignant tumors were frequently associated with pelvic fluid, ascites, or pelvic nodules ($p<0.001$).

Diagnostic performance of adnexal mass prediction models

Table 4 compares the efficacy of different models, namely ResNet50, DenseNet, Vision Transformer, Swin Transformer, and SA, in identifying benign and malignant ovarian tumors (Figure 3). The evaluation metrics used include AUC, sensitivity, specificity, NPV, PPV, Youden index, cutoff value, +LR, -LR, and DOR. The figure depicts the comparison of AUCcurves for different machine learning models. The x-axis represents the false positive rate (FPR), and the y-axis represents the true positive rate (TPR).

Table 1 Histopathological findings in 1555 women with adnexal mass

	Histologic Type	*N*	%
Benign		1196	76.91
	Endometrioid Cystadenoma	420	27.01
	Teratoma	221	14.21
	Serous Cystadenoma	192	12.35
	Mucinous Cystadenoma	78	5.02
	Fibroma and Related Tumors	71	4.57
	Simple Cyst	67	4.31
	Mesosalpinx cyst	54	3.47
	Salpingitis	47	3.02
	Fibrothecoma	19	1.22
	Paraovarian Cyst	5	0.32
	Sertoli-Leydig Cell Tumor (High Grade)	5	0.32
	Benign Brenner Tumor	4	0.26
	Seromucinous Cystadenoma	4	0.26
	Other ovarian benign lesion	9	0.58
Borderline ovarian tumor		53	3.41
	Serous	28	1.80
	Mucinous	21	1.35
	Endometrioid	2	0.13
	Brenner Tumor	2	0.13
Primary ovarian malignant		252	16.21
	Serous Adenocarcinoma	167	10.74
	Clear Cell Carcinoma	30	1.93
	Endometrioid Adenocarcinoma	25	1.61
	Mucinous Adenocarcinoma	9	0.58
	Granulosa Cell Tumor	9	0.58
	Carcinosarcoma	1	0.06
	Sarcoma	3	0.19
	Neuroendocrine Carcinoma	1	0.06
	Sertoli-Leydig Cell Tumor (Low Grade)	1	0.06
	Fibrosarcoma	1	0.06
	Malignant Teratoma	3	0.19
	Dysgerminoma	1	0.06
	Yolk sac tumor	1	0.06
Ovarian metastasis		54	3.47

Table 2 Demographic and Clinical Characteristics of patients with benign and malignant ovarian tumors ($n=1555$)

Characteristic	Benign ($n=1196$)	Malignant ($n=359$)	*P* value
Age (years)	41.0 (32.0–55.0)	54.0 (43.0–64.0)	< 0.001
Menopausal Status	Pre/Post (845/351)	Pre/Post (159/200)	< 0.001
CA125 (U/mL)	17.6 (10.1–40.0)	122.2 (23.4-791.5)	< 0.001
Maximum lesion diameter (mm)*	55.0 (39.0–76.0)	74.0 (46.0-115.0)	< 0.001
Solid Component			
No. of solid components	124 (10.4)	165 (46.0)	< 0.001
Maximum largest solid component diameter (mm)*	24.0 (12.0–39.0)	50.0 (33.0–78.0)	< 0.001
Color Doppler score			
No flow, score 1	725 (60.6)	38 (10.6)	
Minimal flow, score 2	322 (26.9)	60 (16.7)	
Moderate flow, score 3	79 (6.6)	57 (15.9)	
Very strong flow, score 4	70 (5.9)	204 (56.8)	
External Contour			
Regular	187 (59.7)	134 (41.9)	< 0.001
Irregular	126 (40.3)	186 (58.1)	< 0.001
Internal Wall			
Smooth	546 (50.9)	17 (8.8)	< 0.001
Irregular	526 (49.1)	177 (91.2)	< 0.001
Ascites	12 (1.0)	96 (26.7)	< 0.001
Pelvic Nodules	16 (1.3)	78 (21.7)	< 0.001

Among these models, ResNet50, DenseNet, Swin Transformer, and SA achieved high AUC values of 0.91, 0.91, 0.92, and 0.97, respectively. Vision Transformer had a slightly lower AUC of 0.87. In terms of sensitivity, Swin Transformer and SA performed the best sensitivity scores, with values of 87.2% for both models. Specificity was highest for SA at 98.4%, followed by Swin Transformer at 94.3%. Vision Transformer had the lowest specificity at 81.2%.

When considering the NPV, all models performed similarly well, with values above 99.6%. However, there were notable differences in PPV. SA had the highest PPV at 52.0%, while Vision Transformer had the lowest at 8.4%. The Youden index, a measure of overall diagnostic performance, was highest for SA at 0.86. Cutoff values were determined for each model, with values ranging from >0.17 to >3. Additionally, +LR values ranged from 4.49 to 53.18, while -LR values ranged from 0.13 to 0.25. The DOR was highest for SA at 409.08.

Table 5 further compares the efficacy of models in identifying benign and malignant ovarian tumors, with and without the use of CA125, a biomarker for ovarian cancer (Figure 4). The evaluation metrics used are similar to those in Table 4. The results showed that the addition of CA125 did not significantly improve the performance of the models in terms of AUC and sensitivity. However, there were slight improvements in PPV and DOR when CA125 was incorporated. Overall, the performance of the models remained consistent regardless of the presence of CA125.

Channel attention visualization analysis

As illustrated in Fig. 5, the gradient-weighted class activation map are generated by using the gradients of the classification score with respect to the final convolutional feature map. In the Grad-CAM image, the activated (red) area is strongly considered in predicting final results, whereas the blue area is generally not considered in the final result. These findings were compared with justifications provided by clinicians. In cases where the diagnosis was correct, both the models and clinicians focused

Table 3 Demographic and Clinical Characteristics of patients in training set, validation set and test set ($n = 1555$)

Characteristic	Training Set ($n = 1241$)	Validation Set ($n = 154$)	Test Set ($n = 160$)	*P* value
Benign/ Malignant	956/285	119/35	121/39	0.918
Age (years)	44.0 (33.0–57.0)	45.0 (34.0–58.0)	46.0 (33.8–63.0)	0.134
Menopausal Status (Pre/Post)	812/429	99/55	93/67	0.191
CA125 (U/mL)	22.0 (10.9–68.3)	20.8 (11.0–70.0)	19.2 (11.1–51.0)	0.484
Maximum lesion diameter (mm)*	58.0 (40.0–84.0)	53.0 (38.0-78.8)	56.5 (41.0-77.2)	0.805
Solid Component				
No. of solid components	225 (18.1)	29 (18.8)	35 (21.9)	0.172
Maximum largest solid component diameter (mm)*	36.5 (18.0–62.0)	34.0 (19.5–53.0)	38.5 (20.2–57.8)	0.963
Color Doppler score				0.227
No flow, score 1	594 (47.9)	83 (53.9)	86 (53.8)	
Minimal flow, score 2	311 (25.1)	36 (23.4)	35 (21.9)	
Moderate flow, score 3	118 (9.5)	11 (7.1)	7 (4.4)	
Very strong flow, score 4	218 (17.6)	24 (15.6)	32 (20.0)	
External Contour				0.588
Regular	291 (54.0)	35 (58.3)	35 (54.7)	
Irregular	248 (46.0)	25 (41.7)	29 (45.3)	
Internal Wall				0.208
Smooth	425 (41.8)	60 (48.0)	58(46.4)	
Irregular	591 (58.2)	65 (52.0)	67 (53.6)	
Ascites	91 (7.3)	8 (5.2)	9 (5.6)	0.484
Pelvic Nodules	80 (6.4)	7 (4.5)	7 (4.4)	0.417

Table 4 Comparison of the efficacy of ResNet, DenseNet, Vision Transformer, Swin Transformer and SA in identifying benign and malignant ovarian tumors

Model	AUC	Sensitivity	Specificity	NPV	PPV	Youden index	Cutoff	+LR	-LR	DOR
ResNet	0.91 (0.85 - 0.95)	82.1 (60.7 - 88.9)	93.4 (87.5 - 97.1)	99.6 (99.2 - 99.8)	20.3 (7.2 - 45.7)	0.75	>0.58	11.73	0.25	46.92
DenseNet	0.91 (0.86 - 0.95)	84.6 (69.5 - 94.1)	92.6 (86.5 - 96.6)	99.7 (99.3 - 99.8)	26.0 (8.1 - 58.4)	0.77	>0.25	11.47	0.17	67.47
Vision Transformer	0.87 (0.81 - 0.92)	84.6 (69.5 - 94.1)	81.2 (73.1 - 87.7)	99.6 (99.2 - 99.8)	8.4 (4.5 - 15.1)	0.66	>0.17	4.49	0.19	23.63
Swin Transformer	0.92 (0.87 - 0.96)	87.2 (72.6 - 95.7)	94.3 (88.5 - 97.7)	99.7 (99.4 - 99.9)	23.7 (7.9 - 52.7)	0.81	>0.33	15.19	0.14	108.5
SA	0.97 (0.93 - 0.99)	87.2 (72.6 - 95.7)	98.4 (94.2 - 99.8)	99.7 (99.4 - 99.9)	52.0 (8.7 - 92.5)	0.86	>3	53.18	0.13	409.08

on the same regions of interest. Nonetheless, there were instances where both clinicians and DCNNs made incorrect diagnoses. We also compared the areas of interest identified by advanced Sonographers and machine learning models.

We further analyzed six misdiagnosis cases as shown in Fig. 6. Case A was benign, but all four machine learning models predicted it as malignant. The postoperative pathology revealed it to be an endometriotic cyst with old hemorrhage and coffee-colored material, without nodules or papillary growth. The machine learning algorithms may have misinterpreted the old blood clot as a papillary or solid component, erroneously considering it a malignant feature. In Case B, despite being benign, DenseNet, Swin, and Vision Transformer models predicted it as malignant. The postoperative pathology confirmed it to be an endometriotic cyst. However, it differed from typical ground-glass appearance on ultrasound, showing uniform hyperechoic content within the cyst. Analyzing the class activation maps, we observed that the misjudgment models excessively focused on the hyperechoic area, potentially leading to misclassification.

Similarly, in Case C, which was a scenario like Case A with an endometriotic cyst and old hemorrhage, the presence of bleeding clots resembling papillary projections resulted in misdiagnosis by two Transformer models. Case D involved pathological changes due to torsion of an adnexal cyst. Except for the DenseNet model, all other models incorrectly classified it as malignant. This may be attributed to the large size of the tumor, causing the models to miss capturing benign features accurately, leading to misclassification. Additionally, the extensive

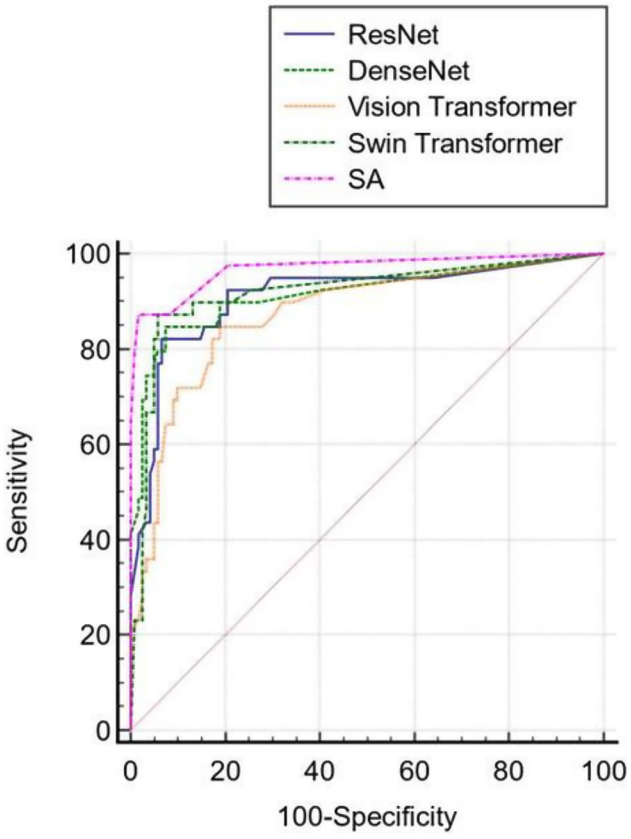

Fig. 3 Comparison of the efficacy of ResNet, DenseNet, Vision Transformer, Swin Transformer and SA in identifying benign and malignant ovarian tumors

Table 5 Comparison of the efficacy of ResNet, DenseNet, Vision Transformer and Swin Transformer in identifying benign and malignant ovarian tumors with or without CA125

	AUC	Sensitivity	Specificity	NPV	PPV	Youden index	Cutoff value	DOR	*P* value
ResNet	0.91(0.85–0.95	82.1(60.7–88.9)	93.4(87.5–97.1)	99.6(99.2–99.8)	20.3(7.2–45.7)	0.75	> 0.58	46.92	
ResNet + CA125	0.90(0.84–0.94)	82.1(66.5–92.5)	93.4(87.5–97.1)	99.6(99.2–99.8)	20.3(7.2–45.7)	0.75	> 0.38	65.84	0.29
DenseNet	0.91(0.86–0.95)	84.6(69.5–94.1)	92.6(86.5–96.6)	99.7(99.3–99.8)	26.0(8.1–58.4)	0.77	> 0.25	67.47	
DenseNet + CA125	0.91(0.85–0.95)	84.6(69.5–94.1)	95.9(90.7–98.7)	99.7(99.3–99.8)	29.6(8.4–65.9)	0.81	> 0.18	129.06	0.53
Vision Transformer	0.87(0.81–0.92)	84.6(69.5–94.1)	81.2(73.1–87.7)	99.6(99.2–99.8)	8.4(4.5–15.1)	0.66	> 0.17	23.63	
Vision Transformer + CA125	0.87(0.81–0.92)	84.6(69.5–94.1)	79.5(71.3–86.3)	99.6(99.2–99.8)	7.8(4.3–13.7)	0.64	> 0.11	21.74	0.71
Swin Transformer	0.92(0.87–0.96)	87.2(72.6–95.7)	94.3(88.5–97.7)	99.7(99.4–99.9)	23.7(7.9–52.7)	0.81	> 0.33	108.50	
Swin Transformer + CA125	0.93(0.88–0.97)	87.2(72.6–95.7)	94.3(88.5–97.7)	99.7(99.4–99.9)	23.7(7.9–52.7)	0.81	> 0.25	108.50	0.23
SA	0.97(0.93–0.99)	87.2(72.6–95.7)	98.4(94.2–99.8)	99.7(99.4–99.9)	52.0(8.7–92.5)	0.86	> 3	409.08	

hemorrhagic necrosis resulting from a 1080° torsion might have caused the models to overly focus on certain benign features and erroneously consider them malignant. Cases E and F were both mature cystic teratomas with neural glial components—a unique subtype of teratomas. Benign teratomas often exhibit characteristic ultrasonographic features, such as mixed echogenicity/white ball and stripes/shadowing [27]. However, these two cases presented with similar solid components and/or thick septations.

The models may have mistakenly classified them as malignant characteristics, potentially resulting in misdiagnosis.

Discussion

This study compared the diagnostic performance of various deep learning models in predicting the malignancy of adnexal masses on ultrasound images. Overall, all four models demonstrated promising results. AUC varies from 0.87 to 0.92. Different models have trade-offs between sensitivity, specificity, positive predictive value, negative predictive value, and positive/negative

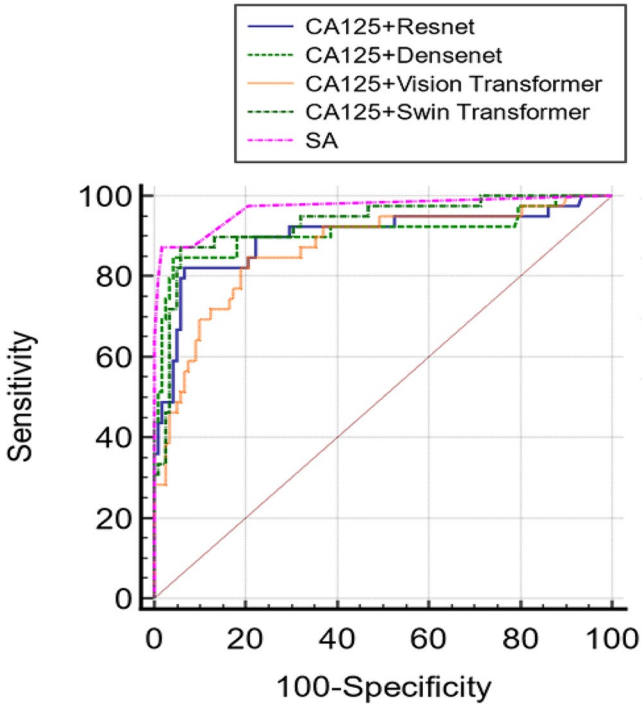

Fig. 4 Comparison of the AUC of ResNet + CA125, DenseNet + CA125, Vision Transformer + CA125, Swin Transformer + CA125 and SA in identifying benign and malignant ovarian tumors

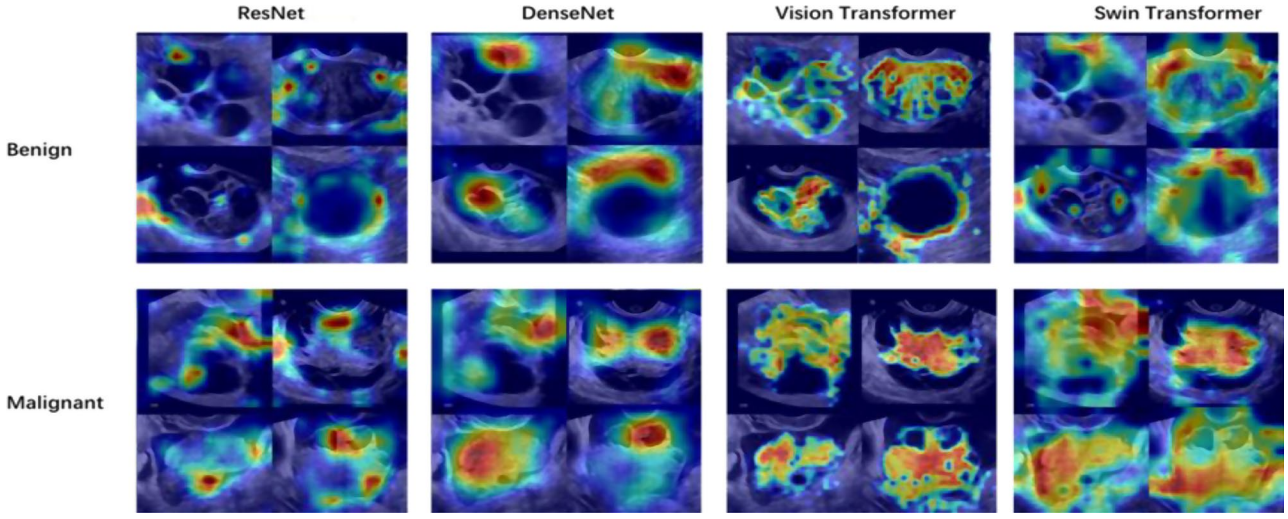

Fig. 5 Visualization of channel attention module

likelihood ratios. However, the Swin Transformer model demonstrated superior diagnostic performance in predicting malignancy in adnexal masses on ultrasound images. It achieved the highest overall accuracy, with a sensitivity of 87.2%, specificity of 94.3%, and an impressive AUC of 0.92, comparable to that of the expert. These superior results can be attributed to the unique features and capabilities of the Swin Transformer model. The Swin Transformer backbone employs shifted windows to extract features at five different scales for self-attention computation. Afterward, a feature pyramid network (FPN) is employed to merge the features from multiple scales. Lastly, a detection head is utilized to predict bounding boxes and their corresponding confidence scores [28].

Previously, most machine learning models used for assisting medical image diagnosis in the field of healthcare have been predominantly CNN-based, such as ResNet and DenseNet. Recently, swin Transformer has demonstrated promising results in applications in

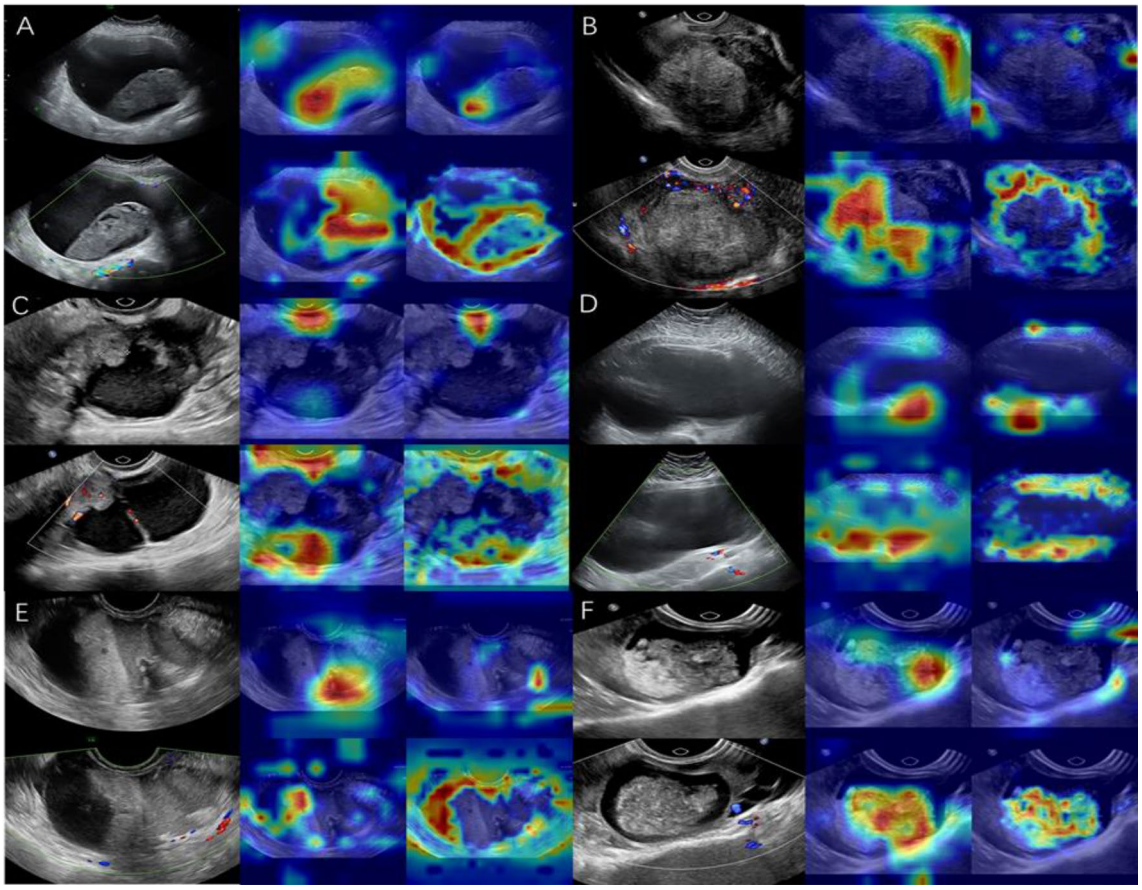

Fig. 6 CAM analysis of 6 cases **(A-F).** The grayscale ultrasound images are shown on the top left, while the Doppler ultrasound images are shown below. On the right side, clockwise from top left, are DenseNet, ResNet, Swin, and VisionTransformer

medical imaging such as differential diagnosis of thyroid nodule, and automated classification of cervical lymph-node-level from ultrasound [29, 30]. However, the fields for assisting ovarian tumor ultrasound diagnosis mainly relied on CNNs, and the use of Swin Transformer model was not reported. ResNet and DenseNet have shown impressive performance in various tasks involving adnexal mass ultrasound image analysis. However, they suffer from the limitation of capturing long-range contextual dependencies due to the restricted receptive field of convolutional layers. In contrast, Transformer networks, including the Swin Transformer mentioned earlier, excel at capturing long-range contextual information. Transformers employ self-attention mechanisms to model the relationships among different positions within an input sequence or image, enabling them to capture both local and long-range dependencies more effectively. By leveraging self-attention, Transformer networks can aggregate information from different parts of an image and capture global contextual dependencies. This paper represents the first attempt to utilize Swin Transformer in this context, and it has achieved favorable diagnostic outcomes.

It is worth noting that the inclusion of CA125 in the models did not significantly improve the diagnostic performance, which aligns with the findings of previous studies [31]. This can be attributed to various factors, including the correlation between tumor markers and certain imaging features leading to information redundancy, insufficient data volume, reactive elevation of CA125 in benign adnexal tumors, and CA125's primary indication of epithelial cell-related pathologies. When developing medical imaging diagnostic models, it is essential to consider these factors, integrate multiple sources of information, and utilize complementary clinical and imaging features to improve accuracy and performance.

The utilization of Grad-CAM has provided valuable insights into the decision-making process of the models by generating class activation maps. These maps effectively highlight the regions in the image that exert the greatest influence on the classification decision. It has been observed that malignant tumors consistently exhibit a higher concentration of red pixels in key areas, such as the solid component. Conversely, benign tumors tend to have a greater number of blue pixels, suggesting

a potential lack of distinct features for benign cases. Through the analysis of six cases, it was determined that the models perform well in identifying common tumor types. However, challenges arise when dealing with specific tumor types, such as mature cystic teratomas with neuronal glial components, or tumors presenting unusual characteristics like endometriotic cysts with hemorrhage. Inaccurate identification of certain tumor characteristics, such as misclassifying old hemorrhagic lesions as solid components, can result in misjudgment and potential misdiagnosis.

To enhance the diagnostic efficacy of the models, additional training data that includes a diverse range of rare and unique tumor cases should be incorporated. By exposing the models to a wider variety of tumor characteristics and presentations, they can acquire a more comprehensive understanding and improve their ability to accurately diagnose such challenging cases. Continued research and refinement of the models can lead to enhanced diagnostic performance and facilitate more accurate identification of rare and complex tumor types.

This study has several strengths. Firstly, a large number of patients were included, which allowed for a robust validation of the transformer model's diagnostic accuracy in ovarian cancer diagnosis. The study also utilized a comprehensive dataset and analyzed a significant number of ultrasound images, contributing to the reliability of the findings. Moreover, the study adhered to strict evaluation protocols based on the IOTA consensus statement, ensuring standardized and consistent assessment of tumor morphology in the ultrasound images. Furthermore, CA125 levels were measured using the same methodology for all patients, increasing the study's reliability. However, the study was conducted at a single center retrospectively, which introduces potential bias in terms of sample distribution and specific patient characteristics.

Overall, this study demonstrates the potential of deep learning models, especially transformer models to accurately predict the malignancy of adnexal masses on ultrasound images.

Abbreviations

ADNEX	Assessment of Different NEoplasias in the adnexa
AUC	Area Under the Curve
CA125	Cancer Antigen 125
CI	Confidence Interval
CNN	Convolutional Neural Network
DenseNet	Densely Connected Convolutional Network
DL	Deep Learning
FPN	Feature Pyramid Network
GCP	Good Clinical Practice
GPU	Graphics Processing Unit
HE4	Human Epididymis Protein 4
IOTA	International Ovarian Tumor Analysis
ML	Machine Learning
NLP	Natural Language Processing
NPV	Negative Predictive Value
OC	Ovarian Cancer
O-RADS	Ovarian-Adnexal Reporting and Data System
PPV	Positive Predictive Value
ResNet	Residual Network
ROMA	Risk of Ovarian Malignancy Algorithm
SA	Subjective Assessment
SR	Simple Rules
Swin Transformer	Shifted Windows Transformer (a type of deep learning model)
TVUS	Transvaginal Ultrasound
US	Ultrasound
ViT	Vision Transformer

Author contributions

CH and FWW conceptualized and designed the study, supervised data collection and reviewed and revised the manuscript. HX collected data, carried out the initial analyses, drafted the initial manuscript, and revised the manuscript. BXH processed the data, utilized machine learning models, and revised the manuscript for machine learning models content. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Funding

Sponsored by Medical Innovation Project of Shanghai Science and Technology Commission (20Y11914000), National Natural Science Foundation of China (grant number 82172601), Natural Science Foundation of Shanghai Science and Technology Commission (20ZR1433700).

Data availability

The images used in this study are available from the corresponding author upon request. All data analyzed in this study are included in the published article.

Declarations

Competing interests

The authors declare no competing interests.

Received: 21 June 2024 / Accepted: 23 October 2024
Published online: 06 November 2024

References

1. Lim MC, Chang SJ, Park B, Yoo HJ, Yoo CW, Nam BH, et al. Survival after Hyperthermic Intraperitoneal Chemotherapy and primary or interval cytoreductive surgery in ovarian Cancer: a Randomized Clinical Trial. JAMA Surg. 2022;157(5):374–83.
2. Kuroki L, Guntupalli SR. Treatment of epithelial ovarian cancer. BMJ (Clinical Res ed). 2020;371:m3773.
3. 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.
4. Brons PE, Nieuwenhuyzen-de Boer GM, Ramakers C, Willemsen S, Kengsakul M, van Beekhuizen HJ. Preoperative Cancer Antigen 125 Level as Predictor for Complete Cytoreduction in Ovarian Cancer: A Prospective Cohort Study and Systematic Review. Cancers. 2022;14(23).
5. Cramer DW, Vitonis AF, Sasamoto N, Yamamoto H, Fichorova RN. Epidemiologic and biologic correlates of serum HE4 and CA125 in women from the National Health and Nutritional Survey (NHANES). Gynecol Oncol. 2021;161(1):282–90.
6. Carvalho JP, Moretti-Marques R, Filho A. Adnexal mass: diagnosis and management. Revista brasileira de ginecologia e obstetricia: revista da Federacao Brasileira das Sociedades de. Ginecol e Obstet. 2020;42(7):438–43.
7. Tavoraitė I, Kronlachner L, Opolskienė G, Bartkevičienė D. Ultrasound Assessment of Adnexal Pathology: Standardized Methods and Different Levels of Experience. Med (Kaunas Lithuania). 2021;57(7).
8. Timmerman D, Van Calster B, Testa AC, Guerriero S, Fischerova D, Lissoni AA, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the

IOTA group. Ultrasound Obstet gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2010;36(2):226–34.
9. Timmerman D, Ameye L, Fischerova D, Epstein E, Melis GB, Guerriero S, et al. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. BMJ (Clinical Res ed). 2010;341:c6839.
10. Van Calster B, Van Hoorde K, Valentin L, Testa AC, Fischerova D, Van Holsbeke C, 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 (Clinical Res ed). 2014;349:g5920.
11. Valentin L, Ameye L, Savelli L, Fruscio R, Leone FP, Czekierdowski A, et al. Adnexal masses difficult to classify as benign or malignant using subjective assessment of gray-scale and Doppler ultrasound findings: logistic regression models do not help. Ultrasound Obstet gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2011;38(4):456–65.
12. López-Úbeda P, Martín-Noguerol T, Luna A. Radiology, explicability and AI: closing the gap. Eur Radiol. 2023;33(12):9466–8.
13. Christiansen F, Epstein EL, Smedberg E, Åkerlund 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 gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2021;57(1):155–63.
14. Gao Y, Zeng S, Xu X, Li H, Yao S, Song K, 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):e179–87.
15. Chen H, Yang BW, Qian L, Meng YS, Bai XH, Hong XW, et al. Deep Learning Prediction of Ovarian Malignancy at US Compared with O-RADS and Expert Assessment. Radiology. 2022;304(1):106–13.
16. Parvaiz A, Khalid M, Zafar R, Ameer H, Ali M, Fraz M. Vision Transformers in Medical Computer Vision -- A Contemplative Retrospection2022.
17. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2000;16(5):500–5.
18. Meys EMJ, Jeelof LS, Achten NMJ, Slangen BFM, Lambrechts S, Kruitwagen R, Van Gorp T. Estimating risk of malignancy in adnexal masses: external validation of the ADNEX model and comparison with other frequently used ultrasound methods. Ultrasound Obstet gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2017;49(6):784–92.
19. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015:770-8.
20. Huang G, Liu Z, van der Maaten L, Weinberger K. Densely Connected Convolutional Networks2017.
21. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv. 2020;abs/2010.11929.
22. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:9992–10002.
23. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vision. 2016;128:336–59.
24. Meinhold-Heerlein I, Fotopoulou C, Harter P, Kurzeder C, Mustea A, Wimberger P, Hauptmann S, Sehouli J. The new WHO classification of ovarian, fallopian tube, and primary peritoneal cancer and its clinical implications. Arch Gynecol Obstet. 2016;293(4):695–700.
25. Prat J. Staging classification for cancer of the ovary, fallopian tube, and peritoneum. Int J Gynaecol Obstet. 2014;124(1):1–5.
26. Piovano E, Cavallero C, Fuso L, Viora E, Ferrero A, Gregori G, et al. Diagnostic accuracy and cost-effectiveness of different strategies to triage women with adnexal masses: a prospective study. Ultrasound Obstet gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2017;50(3):395–403.
27. Timmerman D, Planchamp F, Bourne T, Landolfo C, du Bois A, Chiva L, et al. ESGO/ISUOG/IOTA/ESGE Consensus Statement on preoperative diagnosis of ovarian tumors. Ultrasound Obstet gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2021;58(1):148–68.
28. Tian Y, Zhu J, Zhang L, Mou L, Zhu X, Shi Y et al. Swin Transformer-Based Model for Thyroid Nodule Detection in Ultrasound Images. J visualized experiments: JoVE. 2023(194).
29. Liu Y, Zhao J, Luo Q, Shen C, Wang R, Ding X. Automated classification of cervical lymph-node-level from ultrasound using Depthwise Separable Convolutional Swin Transformer. Comput Biol Med. 2022;148:105821.
30. Chen F, Han H, Wan P, Liao H, Liu C, Zhang D. Joint Segmentation and Differential Diagnosis of Thyroid Nodule in Contrast-Enhanced Ultrasound Images. IEEE Trans Bio Med Eng. 2023;70(9):2722–32.
31. Chen H, Qian L, Jiang M, Du Q, Yuan F, Feng W. Performance of IOTA ADNEX model in evaluating adnexal masses in a gynecological oncology center in China. Ultrasound Obstet gynecology: official J Int Soc Ultrasound Obstet Gynecol. 2019;54(6):815–22.

Publisher's note