A Deep Learning-Based Automatic Recognition Model for Polycystic Ovary Ultrasound Images

Baihua Zhao[1,2], Lieming Wen[2], Yunxia Huang[3], Yaqian Fu[4], Shan Zhou[4], Jieyu Liu[2], Minghui Liu[2], Yingjia Li[1]

[1]Department of Ultrasound, Nanfang Hospital, Southern Medical University, Guangdong, China
[2]Department of Ultrasound Diagnosis, The Second Xiangya Hospital, Central South University, Hunan, China
[3]Department of Ultrasound, The Third Xiangya Hospital, Central South University, Hunan, China
[4]Health Management Center, The Second Xiangya Hospital, Central South University, Hunan, China

Background: Polycystic ovary syndrome (PCOS) has a significant impact on endocrine metabolism, reproductive function, and mental health in women of reproductive age. Ultrasound remains an essential diagnostic tool for PCOS, particularly in individuals presenting with oligomenorrhea or ovulatory dysfunction accompanied by polycystic ovaries, as well as hyperandrogenism associated with polycystic ovaries. However, the accuracy of ultrasound in identifying polycystic ovarian morphology remains variable.

Aims: To develop a deep learning model capable of rapidly and accurately identifying PCOS using ovarian ultrasound images.

Study Design: Prospective diagnostic accuracy study.

Methods: This prospective study included data from 1,751 women with suspected PCOS who presented at two affiliated hospitals at Central South University, with clinical and ultrasound information collected and archived. Patients from center 1 were randomly divided into a training set and an internal validation set in a 7:3 ratio, while patients from center 2 served as the external validation set. Using the YOLOv11 deep learning framework, an automated recognition model for ovarian ultrasound images in PCOS cases was constructed, and its diagnostic performance was evaluated.

Results: Ultrasound images from 933 patients (781 from center 1 and 152 from center 2) were analyzed. The mean average precision of the YOLOv11 model in detecting the target ovary was 95.7%, 97.6%, and 97.8% for the training, internal validation, and external validation sets, respectively. For diagnostic classification, the model achieved an F1 score of 95.0% in the training set and 96.9% in both validation sets. The area under the curve values were 0.953, 0.973, and 0.967 for the training, internal validation, and external validation sets respectively. The model also demonstrated significantly faster evaluation of a single ovary compared to clinicians (doctor, 5.0 seconds; model, 0.1 seconds; $p < 0.01$).

Conclusion: The YOLOv11-based automatic recognition model for PCOS ovarian ultrasound images exhibits strong target detection and diagnostic performance. This approach can streamline the follicle counting process in conventional ultrasound and enhance the efficiency and generalizability of ultrasound-based PCOS assessment.

INTRODUCTION

Polycystic ovary syndrome (PCOS) significantly impacts endocrine metabolism, reproductive function, and mental health in women of reproductive age.[1,2] Between 1990 and 2023, the standards for diagnosing and treating PCOS have been continuously revised.[3-9] Due to variations in regional populations and the complex nature of the disorder[10], diagnosing and managing PCOS remains a persistent challenge in gynecology and endocrinology.

Chinese gynecological experts have reported that as many as 70% of PCOS cases go undiagnosed, and over one-third of patients experience delayed diagnoses.[8] The most recent international and Chinese PCOS guidelines[6,8,9] emphasize the importance of performing ultrasound evaluations of polycystic ovarian morphology (PCOM) in suspected PCOS cases that do not represent both oligomenorrhea or ovulatory dysfunction (O) and hyperandrogenism (H) simultaneously. A large-scale investigation into fertility patterns among Chinese women.[11-14] found that in national epidemiological surveys conducted in 2010

Corresponding author: Yingjia Li, Department of Ultrasound, Nanfang Hospital, Southern Medical University, Guangdong, China

e-mail: lyjia@smu.edu.cn

Received: May 23, 2025 **Accepted:** July 30, 2025 **Available Online Date:** • **DOI:** 10.4274/balkanmedj.galenos.2025.2025-5-114

Available at www.balkanmedicaljournal.org

ORCID iDs of the authors: B.Z. 0000-0001-7185-292X; L.W. 0000-0003-3283-888X; Y.H. 0009-0002-7909-7900; Y.F. 0000-0003-3683-304X; S.Z. 0000-0001-5795-8277; J.L. 0000-0003-4203-9985; M.L. 0000-0002-6548-3264; Y.L. 0009-0004-1206-8608.

Cite this article as: Zhao B, Wen L, Huang Y, Fu Y, Zhou S, Liu J, Liu M, Li Y. A Deep Learning-Based Automatic Recognition Model for Polycystic Ovary Ultrasound Images. Balkan Med J.;

and 2020, about 52% of PCOS patients in China had either the O + PCOM or H + PCOM subtypes[11,12], compared to approximately 30.2% in Europe and the United States of America (USA).[15] The study also indicated that by 2020, the prevalence of PCOS among Chinese women of reproductive age had risen by nearly 65% over the past decade, primarily due to a marked increase in the O + PCOM subtype.[14]

At present, there is no global consensus on the most reliable specificity indicators for the ultrasound-based diagnosis of PCOM[6,15,16], and the diagnostic accuracy among ultrasound practitioners remains variable. The accuracy of follicle counting using 2D versus 3D ultrasound also remains a subject of ongoing debate.[15,16] A meta-analysis Pea et al.[16] demonstrated substantial variation in the diagnostic accuracy of PCOM based on differing PCOS diagnostic criteria and regional populations. While international PCOS guidelines[6,9] recommended the use of ultrasound transducers with frequencies ≥ 8 MHz, Pea et al.[16] found that after stratifying imaging techniques by transducer frequencies < 8 MHz and ≥ 8 MHz, diagnostic accuracy remained unchanged. This may be attributed to the limited consistency in traditional follicle counting methods among ultrasound observers. Pea further reported[16] that 3D ultrasound may not necessarily offer more accurate assessments of PCOS; in fact, it is time-intensive and may lead to an underestimation of follicle count, aligning with findings from Vanden et al.'s[17] study. In addition, the most recent international guidelines[9] formally incorporated serum anti-Müllerian hormone (AMH) levels in defining adult PCOM. However, a standardized threshold for AMH in diagnosing PCOM has not yet been established.[18,19] With the growing integration of artificial intelligence (AI) into medicine, AI has been applied in numerous studies to assist in the diagnosis of PCOS.[20-23] Suha and Islam[22] introduced an enhanced machine learning classification approach to differentiate ultrasound images of PCOS and non-PCOS (NPCOS) cases, achieving an accuracy rate of 99.89%. However, this study introduced both transabdominal and intracavitary ultrasound images, which could introduce classification bias, as current guidelines do not recommend transabdominal ultrasound for evaluating PCOS. To date, related research in China remains limited. Lv et al.[24] proposed a deep learning algorithm for PCOS-assisted detection based on changes observed in the sclera of whole-eye images in females, achieving an accuracy of 92.9%. However, the clinical relevance of these findings is limited.

Globally, most AI research on PCOS is based on small sample sizes, with a lack of large-scale, multicenter studies. Moreover, many of the ultrasound images used in previous studies were sourced from public imaging databases, which varied in imaging techniques and standards, raising concerns about the reliability of the results.

To address this, and to streamline ultrasound-based PCOS evaluation while enhancing its applicability and efficiency, we conducted the first large-scale prospective study among East Asian women to develop a deep learning model capable of rapidly and accurately identifying PCOS ovaries in ultrasound images.

MATERIALS AND METHODS

Ethical approval

This study was approved by the Clinical Research Ethics Committees of the Second Xiangya Hospital (approval number: 2019-036; date: 06.03.2019) and the Third Xiangya Hospital of Central South University (approval number: 2019-066; date: 06.06.2019). All patients provided written informed consent.

Research participants

This prospective study included female patients who attended the gynecology and endocrinology clinics at the Second Xiangya Hospital (center 1) and the Third Xiangya Hospital (center 2) of Central South University between November 2019 and December 2024.

Inclusion criteria were as follows: women of reproductive age (20-44 years) suspected of having PCOS, presenting with menstrual irregularities, infertility, abnormal weight gain or obesity, H (including clinical signs), and an ultrasound diagnosis of PCOM.

Exclusion criteria were as follows: ① use of hormone-containing medications (including oral contraceptives) within the past 3 months; ② pregnancy at the time of screening; ③ coexisting tumors in the uterus, ovaries, or other sites; ④ incomplete clinical information; or ⑤ poor-quality ultrasound images of both ovaries, or the presence of a corpus luteum or follicles measuring 10 mm or more in diameter. Poor-quality ultrasound images refer to those in which the ovarian outline and internal structure could not be clearly identified by the ultrasound physician.

The Rotterdam criteria were used as the diagnostic gold standard for PCOS, as the Chinese PCOS guidelines have consistently been based on these criteria.[4,7,8] The final clinical diagnosis for each patient was recorded.

Clinical and ultrasound data collection

The following patient characteristics were recorded: age, height (cm), weight (kg), body mass index (BMI), menstrual history, reproductive history, signs of H, and history of other diseases. Levels of sex hormones and thyroid hormones were also documented.

The ultrasound machines used at the two centers included SonoScape P60 (SonoScape, Shenzhen, China), Mindray Resona R7 (Mindray Medical, Shenzhen, China), GE E8 (GE Healthcare, Milwaukee, WI, USA), GE E10 (GE Healthcare, Milwaukee, WI, USA), and Voluson S6 (GE Healthcare, Milwaukee, WI, USA). The preset instrument parameters were as follows: intracavitary ultrasound frequency range, 4.0-9.0MHz; imaging depth, 7-8 cm; fan angle range, 20°-180°; speckle suppression, level 4; spatial compounding, level 2; mechanical index, < 1.0; and thermal index, < 1.0.

Intracavitary ultrasound examinations, image acquisition, and ovarian data annotation were conducted by gynecologic ultrasound physicians with over 10 years of experience at each center-three physicians from center 1 and two from center 2. Before the study began, 20 ovarian ultrasound images were randomly selected from each center's ultrasound workstation, comprising 20 PCOS and 20 NPCOS images in total. The physicians responsible for initial image

selection were not involved in subsequent image acquisition or annotation. The five ultrasound doctors diagnosed these 40 ovarian images for PCOM or non-PCOM. Inter-observer agreement among the five physicians was assessed, yielding a kappa value of 0.91 (95% confidence interval, 0.86-0.97). These 40 images used for consistency testing were excluded from the main study.

During data collection, each ovary was diagnosed for PCOM based on the gold standard by the ultrasound physician. Cases with uncertain PCOM diagnoses were documented. If high-quality ultrasound images were available for both ovaries, both were included; if only one ovary had a high-quality image, only that side was retained. For each ovary, the two-dimensional ultrasound image displaying the highest number of follicles was stored. To capture the maximum number of follicles within one section, one to three images per ovary were stored from different angles. If multiple-angle images showed substantial overlap, the redundant overlapping images were excluded. All ultrasound images from the same patient were acquired during the same session, with no variation in imaging conditions during acquisition.

Data structure and partition

Patients from center 1 were assigned to the training and internal validation sets using stratified random sampling at a 7:3 ratio. Patients from center 2 were included in the external validation set. All ovarian images were allocated to the training, internal validation, or external validation sets based on the patient-level assignment described above. No patient and their corresponding images appeared in more than one dataset.

Analysis of image sample similarity and independence

To assess the similarity and independence of the ovarian ultrasound image samples, the grayscale values and histograms of the images were extracted and analyzed. The normalized structural similarity index (SSIM) and mutual information (MI) were calculated for each pair of images to evaluate their degree of similarity and correlation. Statistical differences in grayscale histograms were used to demonstrate the independence of the image samples.

Development of a deep learning model for automatic PCOS recognition

In this study, we utilized YOLOv11, an end-to-end object detection model based on convolutional neural networks, as the modeling framework. The source code for YOLOv11 is publicly available on GitHub at https://github.com/ultralytics/ultralytics. The technical workflow for model development based on YOLOv11 is illustrated in Figure 1.

Preprocessing of modeling data: Initially, all original ovarian ultrasound images in the training set underwent manual or automated preprocessing to enhance data diversity. Subsequently, target-specific annotations-including bounding boxes and category labels-were applied. The preprocessed training data were then divided, using stratified random sampling, into three subsets: a model training set, a model testing set, and a model validation set. The validation set was used to support early stopping and fine-tuning of parameters to prevent model overfitting.

Model architecture: The YOLOv11 architecture is composed of three main components: a backbone network, a neck network, and a head network, which work together to perform object detection. The backbone network extracts fundamental features from the input image and progressively downsamples it through convolutional and pooling layers to generate multiscale feature maps. The neck network integrates and processes these multiscale features, applying operations such as upsampling, downsampling, and feature concatenation to enhance the representation capacity of the features. This allows the model to better capture object characteristics across varying scales before passing them to the head network for prediction. The head network is responsible for outputting the predicted category and the coordinates of the bounding box for each target. It contains two branches: a classification branch, which estimates the probability that the detected object is PCOS or NPCOS, and a regression branch, which predicts the bounding box coordinates of the target.

Defining the loss function: The appropriate loss function was selected based on the nature of the task, specifically the relationship between

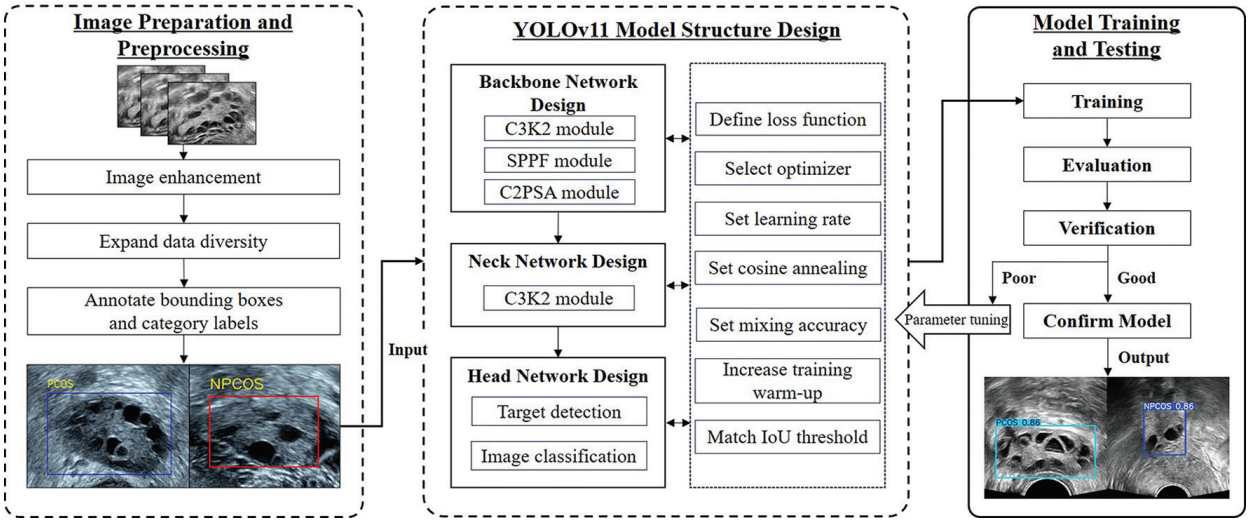

FIG. 1. Technology roadmap for modelling in YOLOv11

the true and predicted categories. In this study, a multitask loss function was employed, which could include classification loss, regression loss, confidence loss, VariFocal loss, and bounding box loss. In addition, each predicted bounding box was matched to its corresponding ground truth using a one-to-one intersection over union (IoU) threshold. When the IoU value between a predicted box and the ground truth box was ≥ 0.7, it was considered a correct prediction.

Optimizer selection and learning rate settings: The AdamW optimizer was chosen for its ability to mitigate overfitting by adaptively adjusting the learning rates of individual parameters. The initial learning rate was set at 0.0001, and a cosine annealing strategy was employed to adjust it during training. Mixed precision training and extended warm-up phases were also used to improve the model's convergence speed.

Model training, testing, and validation

The model was trained to detect the target ovary in ultrasound images and to classify cases as PCOS or NPCOS. The process involved multiple iterations of training, testing, and validation, during which the loss function was calculated and gradients were computed using backpropagation. The optimizer then updated the model parameters accordingly. After training was completed, the ovarian ultrasound images from both the internal and external validation sets were input into the trained model using the same set of parameters for performance evaluation.

Comparison of recognition time between the model and a senior ultrasound physician for PCOM diagnosis

A senior ultrasound physician at center 1 performed intracavitary ultrasound scans on the ovaries of 20 additional patients suspected of having PCOS, recording the time taken to scan each ovary and diagnose PCOM. Correspondingly, the two-dimensional ultrasound image of each ovary containing the highest number of follicles was input into the YOLOv11 model, and the model's recognition time for each ovary was recorded.

Statistical analysis

Statistical analyses were conducted using SPSS version 29.0 and Python version 3.8.3. Inter-observer agreement was evaluated using Fleiss' Kappa. For variables with non-normal distributions, the median and interquartile range were used for description; for variables with normal distributions, the mean and standard deviation were used. The Mann-Whitney U test was applied for comparisons involving continuous variables. A single-sample t-test was used for data with a normal distribution, while the single-sample rank sum test was used for non-normally distributed data ($p < 0.05$ was considered statistically significant). The SSIM ranged from -1 to 1, with a mean value above 0.8 indicating a high degree of overall image similarity. MI ranged from 0 to 1, and a mean value above 0.5 indicated substantial overall image dependency. Histogram differences were analyzed using the chi-squared test of homogeneity, and the chi-squared statistic (chi-squared distance) was calculated. For multiple comparisons involving fewer than 10 tests, the Bonferroni correction was applied to adjust the significance level (original $\alpha = 0.05$, adjusted $\alpha = 0.0167$), with $p < 0.0167$ indicating statistical significance. For more than 10 comparisons, the Benjamini-Hochberg correction was used [false discovery rate (FDR) < 0.05], and results with an FDR-adjusted q-value < 0.05 were considered significant. Model performance was evaluated by calculating accuracy, precision, recall, F1 score, and mean average precision (mAP). A confusion matrix was generated, and the area under the curve (AUC) was computed to evaluate model effectiveness.

Code availability

The model code is accessible at http://github.com/LittleStoneHouse/YOLOv11-.git.

RESULTS

General results

This study collected data from 1,751 female patients with clinically suspected PCOS, comprising 1,451 patients (82.9%) from center 1 and 300 patients (17.1%) from center 2. Of these, 933 patients (53.3%) were ultimately included in the analysis-781 from center 1 and 152 from center 2-while 818 patients (46.7%) were excluded (Figure 2).

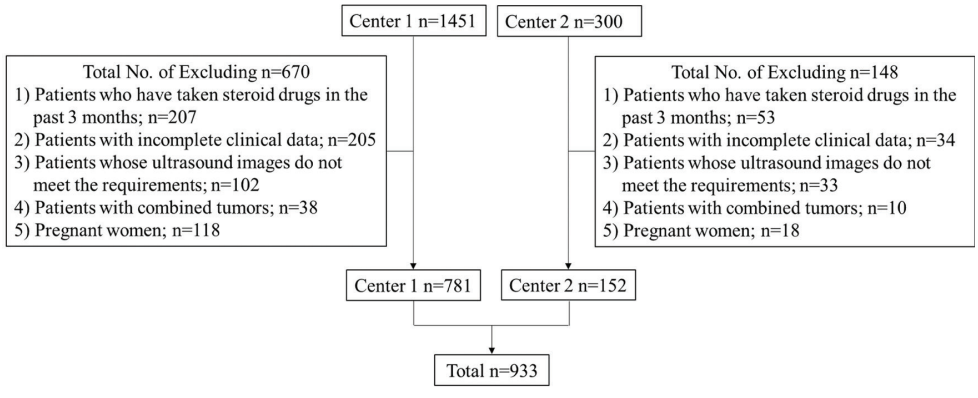

FIG. 2. Inclusion of patients from two centers.

The 933 included patients ranged in age from 20 to 44 years. Their general characteristics are presented in Table 1. No significant differences were observed between the two centers in terms of age, height, weight, BMI, age at menarche, or history of pregnancy and childbirth ($p > 0.05$). However, a significant difference was found in the average number of days in the menstrual cycle between the two centers ($p = 0.030$).

Among the 781 patients from center 1, 365 (46.7%) were diagnosed with PCOS, while 416 (53.3%) were diagnosed with NPCOS. In center 2, 76 out of 152 patients (50.0%) were diagnosed with PCOS, and the remaining 76 (50.0%) were diagnosed with NPCOS. The diagnostic criteria and population distribution details are shown in Figure 3.

Patient and image allocations to the training set, internal validation set, and external validation set

Of the 781 patients enrolled from center 1, 547 (70%) were assigned to the training set and 234 (30%) to the internal validation set through randomization. All 152 patients from center 2 comprised the external validation set (Table 2).

In center 1, 123 patients (29 in the training set and 94 in the internal validation set) contributed only one ovarian ultrasound image each, as the contralateral ovarian images were excluded due to poor quality. A total of 587 patients (448 in the training set and 139 in the internal validation set) had two usable images, and 71 patients (70 in the training set and 1 in the internal validation set) had three images included. In center 2, 92 patients provided only one ovarian ultrasound image, with the corresponding contralateral images excluded due to poor quality, while 60 patients contributed two images (Table 2). Overall, 1,722 ovarian ultrasound images were included, consisting of 1135 images (65.9%) in the training set, 375 images (21.8%) in the internal validation set, and 212 images (12.3%) in the external validation set.

Analysis results of the similarity and independence of the image samples

For patients in both centers with two or three ovarian ultrasound images, the mean SSIM values were significantly below 0.8 ($p < 0.001$), and the mean MI values were significantly below 0.5 ($p < 0.001$), indicating low overall structural similarity and low dependency between each pair of ovarian images. The chi-squared distance of the grayscale histogram was significantly greater than 0 ($p < 0.001$), demonstrating statistical differences in the grayscale pixel distributions of the ovarian images when compared pairwise (Table 3).

TABLE 1. General Characteristics of the Patients in the Two Centers.

Characteristic	Center 1 (n=781)	Center 2 (n=152)	Test value	*p* value
Age (year)	28 (24, 28)	29 (23, 24)	Z = -0.677	0.499
Height (cm)	160 (156, 164)	159 (156, 162)	Z = -1.641	0.101
Weight (kg)	57.5 (53, 63)	57.8 (54, 63)	Z = 0.210	0.984
BMI	22.5 (21.4, 24.0)	22.7 (21.7, 24.2)	Z = 1.583	0.113
Age of menarche (year)	13 (13,13)	13 (12,13)	Z = -1.453	0.146
Average menstrual cycle (days)	39 (30, 48)	35 (30, 45)	Z = -2.168	0.030
Gravidity	0 (0, 2)	0 (0, 2)	Z = 1.147	0.251
Parity	0 (0, 1)	0 (0, 1)	Z = 1.061	0.288

BMI, body mass index = Weight (kg)/Height(m)^2; non-normal distribution variables are described by median and interquartile ranges.

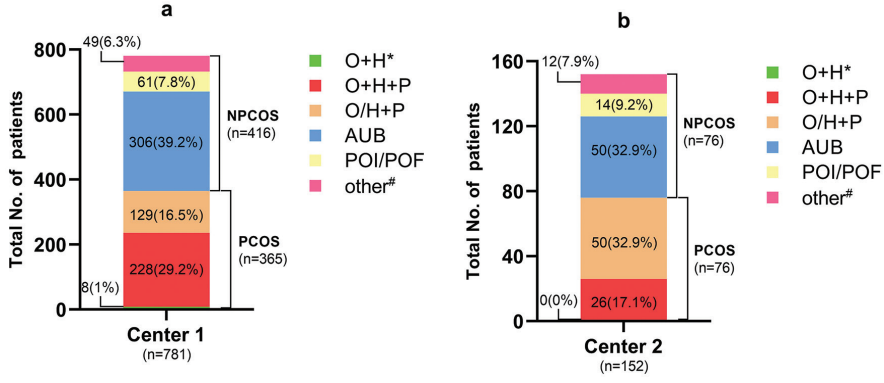

FIG. 3. (a, b) Diagnostic items and population distributions of PCOS and NPCOS patients. The patients in center 1 (a). The patients in center 2 (b).
*PCOS, polycystic ovary syndrome; NPCOS, non-polycystic ovary syndrome; O, oligomenorrhea/ovulation disorder; H, hyperandrogenism; P, polycystic ovarian morphology; AUB, abnormal uterine bleeding; POI, premature ovarian insufficiency; PDF, premature ovarian failure. *Ultrasound doctors determined the absence of PCOM and those whose PCOM could not be fully determined. # Individuals with AUB caused by pituitary, thyroid, adrenal, or kidney diseases.*

TABLE 2. Distribution of the Patients and Ultrasound Images in the Training Set and the Internal and External Validation Sets.

Sets	Diagnosis	No. of patients	No. of patients with 1 image*	No. of patients with 2 images#	No. of patients with 3 images&	Total No. of images
Training (center 1)	PCOS	257	5	184	68	577
	NPCOS	290	24	264	2	558
Internal validation (center 1)	PCOS	108	36	71	1	181
	NPCOS	126	58	68	0	194
External validation (center 2)	PCOS	76	39	37	0	113
	NPCOS	76	53	23	0	99
Total		933	215	647	71	1722

PCOS, polycystic ovary syndrome; NPCOS, non-polycystic ovary syndrome; No., number.
After storing 1-3 images per ovary and removing overlapping and unclear images:
*Only included one image of one ovary (left or right);
#Included one image of each ovary (left and right);
&Included two images in one ovary and one image in the other ovary.

TABLE 3. Results of Similarity and Dependency Analysis for the Images in Two Centers.

Indicators	Center 1		Center 2
	Comparisons of two samples (587 pairs of images)	Comparisons of three samples in pairs (213 pairs of images)	Comparisons of two samples (60 pairs of images)
SSIM	0.48 (0.31, 0.62)*	0.33 (0.22, 0.54)*	0.48 ± 0.14#
MI	0.28 (0.19, 0.34)*	0.19 (0.16, 0.35)*	0.26 (0.18, 0.35)*
Chi^2	2.39 (2.18, 3.28)*	2.45 (1.35, 3.33)*	2.16 (1.11, 3.06)*

SSIM, structural similarity index; MI, mutual information; Chi^2, Chi-square distance. Normal distribution variables are described by mean and standard deviation. Non-normal distribution variables are described by median and interquartile ranges. *: single sample rank sum test, $p < 0.001$; #: single sample t-test, $p < 0.001$.

In pairwise comparisons of image samples, all multiple tests in both centers (587 in center 1 and 60 in center 2) revealed statistically significant differences before and after test level adjustment (FDR threshold = 0.05, $q < 0.05$), confirming substantial differences in the histogram distributions of each pair of ovarian images. The median and interquartile range of Cramér's V effect size were 0.37 (0.33, 0.40) in center 1 and 0.33 (0.31, 0.40) in center 2, respectively.

For comparisons involving three image samples, results from 71 tests in center 1 showed statistically significant differences ($p < 0.05$), indicating that the histogram distributions of the three ovarian images per patient varied significantly. Similarly, 213 pairwise comparisons also showed statistical differences ($p < 0.0167$) before and after correction, supporting the presence of significant differences in histogram distributions. The median to interquartile range of Cramér's V effect size was 0.35 (0.30, 0.39).

Modeling evaluation results of the YOLOv11 model

Throughout training, the values of the model's various loss functions progressively decreased, while the performance metrics such as accuracy, recall, and mAP steadily improved, indicating good model convergence. According to the precision-recall curve, the model achieved an mAP of 95.7% at an IoU threshold of 0.5, with category-specific mAPs of 97.2% for PCOS and 94.1% for NPCOS (Figure 4a). The corresponding diagnostic performance metrics are presented in Table 4.

Performance evaluation results of the YOLOv11 model on internal and external validation sets

The model achieved automatic diagnostic accuracies of 97.3% (365/375) on the internal validation set and 96.7% (205/212) on the external validation set. In the internal validation set, the model reached an mAP of 97.6% at an IoU threshold of 0.5, with mAPs of 97.5% for PCOS and 97.8% for NPCOS (Figure 4b). In the external validation set, the mAP was 97.8% at an IoU threshold of 0.5, with mAPs of 98.5% and 97.1% for PCOS and NPCOS, respectively (Figure 4c). Related diagnostic performance data are detailed in Table 4.

Furthermore, in the internal validation set, 2 patients diagnosed with PCOS showed suspected negative PCOM results, and 16 patients with NPCOS showed suspected PCOM. In the external validation set, five NPCOS patients had suspected PCOM. The model accurately identified the ovarian images for all 23 of these cases.

Error analysis of the YOLOv11 model in the internal and external validation sets

The confusion matrices for both the internal and external validation sets are presented in Figure 5.

In both datasets, the model generated a small number of false positive and false negative results when identifying the target ovary, primarily involving areas of the image background. Detailed analysis indicated that these errors were mainly due to suboptimal image quality-for example, when the ovary's brightness was either too similar to or too different from that of the background.

In terms of classification and diagnosis of the target ovary, there was one missed detection and one misclassification among the 194 NPCOS images in the internal validation set. Within the 181 PCOS images from the same set, one image was not detected, and eight were misclassified. In the external validation set, 3 of the 99 NPCOS images were not recognized as the target, and 3 were misclassified. Among the 113 PCOS images, four were misclassified. The primary factors contributing to classification errors included indistinct ovarian outlines and internal structures, presence of uterine or pelvic blood vessels adjacent to the ovaries, follicle diameters approaching 10 mm, and large central regions of the ovary exhibiting strong stromal echogenicity. Representative examples of these errors are illustrated in Figure 6.

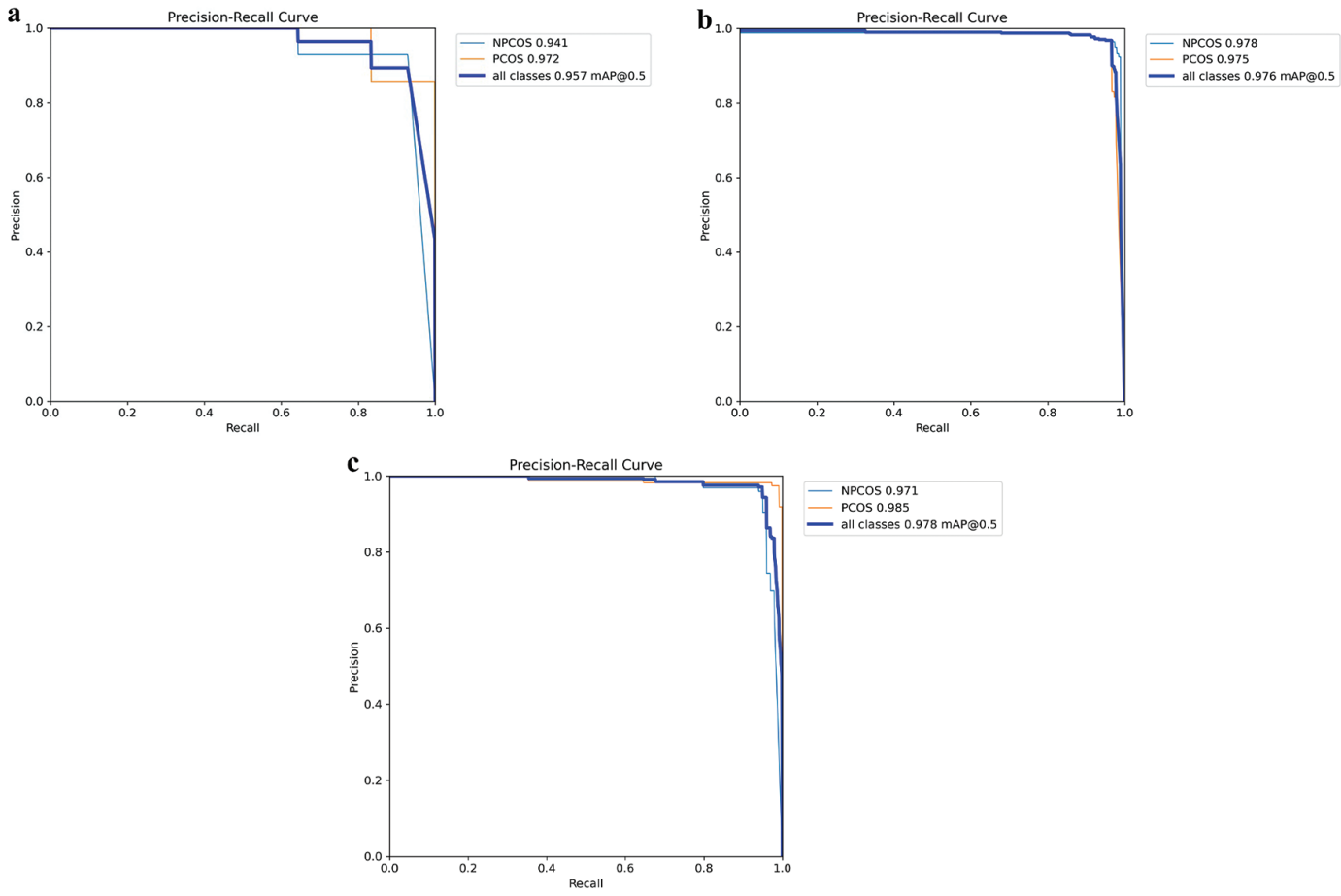

FIG. 4. (a-c) Precision-recall curves of YOLOv11. The precision-recall curve for modelling (a). The precision-recall curve for internal validation (b). The precision-recall curve for external validation (c).

NPCOS, non-polycystic ovarian syndrome; PCOS, polycystic ovarian syndrome; mAP, mean average precision.

TABLE 4. Performance Indicators of the YOLOv11 Model for the Automatic Recognition and Diagnosis of PCOS in the Training Set and the Internal and External Validation Sets.

Sets	Acc %	Rec %	Pre %	F1-score %	mAP	AUC (95% CI)
Training	95.3	95.0	95.0	95.0	95.7	0.953 (0.936-0.969)
Internal validation	97.3	95.0	99.4	96.9	97.6	0.973 (0.953-0.992)
External validation	96.7	96.5	97.3	96.9	97.8	0.967 (0.939-0.995)

PCOS, polycystic ovary syndrome; Acc, accuracy; Rec, recall rate; Pre, precision; mAP, mean average precision; AUC, area under the curve; CI, confidence interval.

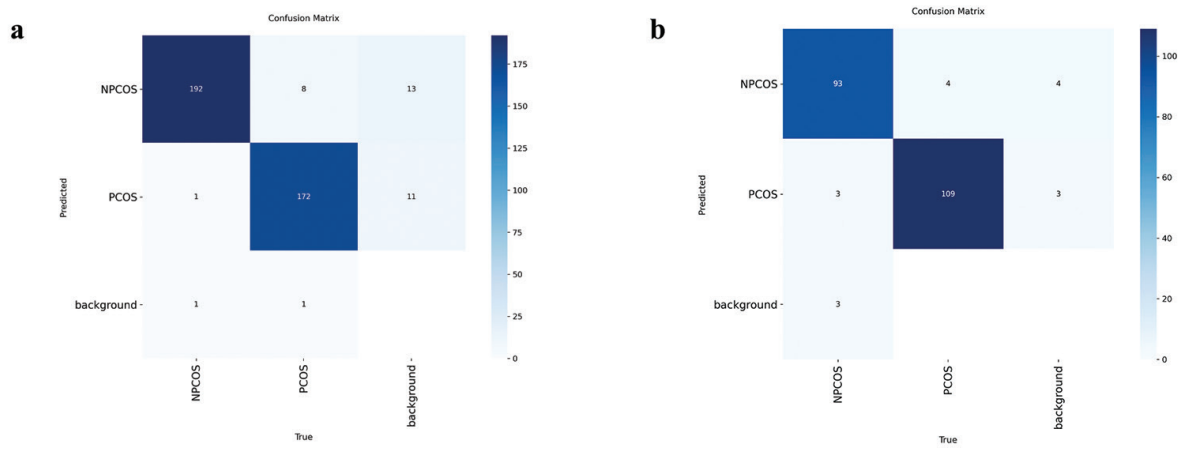

FIG. 5. (a, b) The confusion matrices of the model. The confusion matrix for internal validation (a). The confusion matrix for external validation (b).
NPCOS, non-polycystic ovarian syndrome; PCOS, polycystic ovarian syndrome.

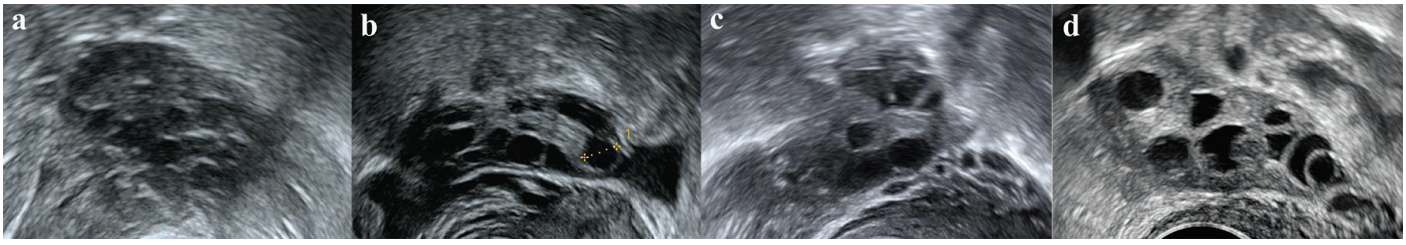

FIG. 6. (a-d) Visual examples of error cases in the validation sets. An ovarian ultrasound image of a PCOS patient with relatively blurry ovarian outlines and internal structures (a). An ovarian ultrasound image of a PCOS patient with uneven image brightness (b). An unrecognized ovarian image of a PCOS patient with uterine and pelvic blood vessels appearing near the ovary (c). An ovarian image with recognition error in a PCOS patient for the follicle diameters close to 10 mm.
PCOS, polycystic ovarian syndrome.

Comparison of recognition and diagnosis time between the model and a senior ultrasound physician

Among 20 patients with suspected PCOS (9 diagnosed with PCOS and 11 with NPCOS), a total of 40 ovarian images were analyzed (18 from PCOS and 22 from NPCOS cases). The median time taken by senior physicians to evaluate each ovary was approximately 5.0 (4.0, 6.0) seconds, while the model required about 0.1 seconds per ovary. The time difference between the two groups was statistically significant ($p < 0.01$).

DISCUSSION

Awareness of the adverse health impacts of PCOS is growing among women. Ultrasound remains the primary imaging modality for gynecological assessments, though its diagnostic accuracy for PCOS can vary. In this study, a deep learning model was developed to analyze ultrasound images and enable rapid and accurate identification of PCOS-related ovarian features.

This study is the first large-scale prospective study using ultrasound images of PCOS in East Asian women, with rigorous inclusion and exclusion criteria applied to samples from two distinct centers. The general clinical characteristics of patients from both centers were comparable. In addition, since the ovaries are symmetrical organs in females, the final analysis included bilateral ovarian ultrasound images. Given this anatomical feature, the structural similarity and dependence between multiple images from the same patient were assessed prior to model training. These similarities and dependencies were found to be low, with no statistically significant correlations between images, thereby ensuring the methodological rigor and reliability of the model's development.

Cheng and Mahalingaiah[25] created and evaluated two machine learning algorithms using 39,093 ultrasound reports from 25,535 women to categorize PCOM, achieving accuracies of 97.6% and 96.1%, respectively. The author emphasized that further work in PCOS automation should focus on direct feature extraction from original ultrasound images. Similarly, Nsugbe[26] developed an AI-based decision support system for early PCOS diagnosis using data from the publicly available Kaggle database. This dataset included 364 NPCOS and 177 PCOS patients, with 41 features spanning metabolic, imaging, hormonal, and biochemical data. Ten machine learning algorithms were compared, and the highest-performing model achieved 93% accuracy.

Our YOLOv11 model effectively enhances the detection of small objects in images using anchor-free technology, making it well-

suited for this study. The ultrasound images were obtained through multiple ultrasound instruments, contributing to the diversity of the dataset. The training process for the model was stable, and the training and validation results were favorable. The model demonstrated high robustness and generalizability.

Both the training set and the internal and external validation sets for the model demonstrated good AUCs (0.953, 0.973, and 0.967, respectively). The recall rates were high in all the sets (95.0%, 95.0%, and 96.5%), indicating that the model had strong sensitivity to PCOS cases and a low likelihood of missed diagnoses. The model obtained high F1 scores in all sets (95.0%, 96.9%, and 96.9%), which indicated a well-maintained balance between accuracy and recall. This demonstrates the model's ability to maintain a high true positive rate and a low false positive rate. In addition, the mAPs of the model were good across all sets (95.7%, 97.6%, and 97.8%), indicating that the model's effectiveness in accurately detecting target location and category recognition, which is particularly important for clinical applications.

Ultrasound images of poor quality were excluded from this study. These were images in which the outline and internal structures of the ovary were not clearly distinguishable to the ultrasound physician. Contributing factors to poor image quality included obesity, ovarian position being far away from the ultrasound transducer due to pelvic adhesions, and unavoidable intestinal gas interference in the patient's pelvic region.

The primary cause of model errors was the relatively low quality of certain images. The error cases shown in the visualization examples may be due attributed the fact that while most ovaries are situated independently within the pelvic cavity, some are closely adjacent to the uterus and pelvic blood vessels, and a few exhibit strong stromal echoes-conditions that are relatively uncommon in the dataset. In addition, based on the Rotterdam Guidelines for PCOS[4], ovaries containing follicles with diameters of 10 mm were excluded. When PCOS ovaries contained follicles approaching 10 mm in diameter, the model occasionally misclassified them. These specific circumstances led to relatively limited exposure during training, which may have contributed to errors in such cases.

Previous studies have indicated that not all PCOS patients exhibit PCOM[15], and around 25% of women with normal reproductive function may show PCOM features[7,19]. Furthermore, PCOM may also be observed in women with coexisting pituitary, thyroid, adrenal, and kidney diseases[7,8]. In our study, such patients accounted for a relatively small proportion of the sample, and the model accurately identified PCOS patients without PCOM and NPCOS patients with suspected PCOM in both the internal and external validation sets. This might be due to the deep learning model's ability to extract more informative features from ovarian images, offer an advantage in recognition. However, more samples may be necessary for further validation.

The YOLOv11 model developed in this study can rapidly and accurately identify the target ovary, operating at a diagnostic speed 50 times faster than that of a senior ultrasound doctor. This clearly highlights the advantages of AI models in medical image recognition and diagnosis. In addition, the model needs only a single ovarian ultrasound section with the highest number of follicles, simplifying the process of follicle counting in PCOS assessments and greatly improving the efficiency.

Among the patients included in our study, gynecological patients accounted for a relatively large proportion of the sample, and fewer women had PCOM with regular menstrual cycles. This resulted in certain bias in sample coverage, and the efficiency of the model should be further validated with samples that allow broader disease coverage. In addition, this study did not include adolescent patients with PCOS, indicating need for further research for this particular population.

In this study, an automatic recognition model for PCOS ovarian ultrasound images was established using the YOLOv11 deep learning framework, demonstrating good target recognition ability and diagnostic efficiency. This method can simplify the conventional follicle counting process based on ultrasound and enhance the universality and speed of PCOS ultrasound evaluation. The model has high potential for clinical application. As the first large-sample prospective study on ultrasound-based evaluation of PCOS in East Asian women at two centers, this study may serve as a foundation for evidence-based medicine for patients with PCOS.

Acknowledgments: The authors thank Hunan Future Smart Healthcare Co., Ltd., for their technical support during the modelling process.

Ethics Committee Approval: This study was approved by the Clinical Research Ethics Committees of the Second Xiangya Hospital (approval number: 2019-036; date: 06.03.2019) and the Third Xiangya Hospital of Central South University (approval number: 2019-066; date: 06.06.2019).

Informed Consent: All patients provided written informed consent.

Data Sharing Statement: The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Authorship Contributions: Concept- B.Z., M.L., Y.L.; Design- B.Z., L.W., Y.L.; Supervision- L.W., M.L., Y.L.; Fundings- B.Z.; Materials- B.Z., Y.H., Y.F., S.Z., J.L.; Data Collection or Processing- B.Z., Y.H., Y.F., S.Z., J.L.; Analysis or Interpretation- B.Z., Y.H., Y.F., S.Z., J.L.; Literature Review- B.Z., J.L.; Writing- B.Z., Y.L.; Critical Review- L.W., M.L., Y.L.

Conflict of Interest: The authors declare that they have no conflict of interest.

Funding: This study was supported by the Natural Science Foundation of Hunan Province (Project No.: 2024JJ9192).

REFERENCES

1. Owens LA, Franks S. Polycystic ovary syndrome: origins and implications: the impact of polycystic ovary syndrome on reproductive health: a narrative review. *Reproduction*. 2025;169:e240485. [CrossRef]
2. Joham AE, Norman RJ, Stener-Victorin E, et al. Polycystic ovary syndrome. *Lancet Diabetes Endocrinol*. 2022;10:668-680. Erratum in: *Lancet Diabetes Endocrinol*. 2022;10:e11. [CrossRef]
3. Zawadski JK, Dunaif A. Diagnostic criteria for polycystic ovary syndrome: towards a rational approach. In: Dunaif A, Givens JR, Haseltine F, editors. Polycystic ovary syndrome. *Boston: Blackwell Scientific*; 1992:377-384. [CrossRef]
4. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril*. 2004;81:19-25. [CrossRef]
5. Azziz R, Carmina E, Dewailly D, et al. Positions statement: criteria for defining polycystic ovary syndrome as a predominantly hyperandrogenic syndrome: an Androgen Excess Society guideline. *J Clin Endocrinol Metab*. 2006;91:4237-4245. [CrossRef]
6. Teede HJ, Misso ML, Costello MF, et al. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Hum Reprod*. 2018;33:1602-1618. Erratum in: *Hum Reprod*. 2019;34:388. [CrossRef]

7. Guideline Expert Group and Endocrinology Group of the Obstetrics and Gynecology Branch of the Chinese Medical Association. Guidelines for diagnosis and treatment of polycystic Ovary in China. *Chin J Obstet Gynecol.* 2018;53:2-6. [CrossRef]
8. Expert Consensus Compilation Group for the Pathway of Diagnosis and Management of Polycystic Ovary Syndrome. Expert consensus on the pathway of diagnosis and management of polycystic ovary syndrome. *Chin J Reprod Contracep.* 2023;43:337-345 [CrossRef]
9. Teede HJ, Tay CT, Laven JJE, et al. Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *J Clin Endocrinol Metab*. 2023;108:2447-2469. [CrossRef]
10. Salari N, Nankali A, Ghanbari A, et al. Global prevalence of polycystic ovary syndrome in women worldwide: a comprehensive systematic review and meta-analysis. *Arch Gynecol Obstet*. 2024;310:1303-1314. [CrossRef]
11. Xue S, Yang G. Progress in diagnosis and therapy of polycystic ovary syndrome. *Chin J Endocrinol Metab.* 2020;36:88-92. [CrossRef]
12. Li R, Zhang Q, Yang D, et al. Prevalence of polycystic ovary syndrome in women in China: a large community-based study. *Hum Reprod.* 2013;28:2562-2569. [CrossRef]
13. Zhou Z, Zheng D, Wu H, et al. Epidemiology of infertility in China: a population-based study. *BJOG.* 2018;125:432-441. [CrossRef]
14. Yang R, Li Q, Zhou Z, et al. Changes in the prevalence of polycystic ovary syndrome in China over the past decade. *Lancet Reg Health West Pac*. 2022;25:100494. [CrossRef]
15. Pace L, Waldeck J, Chan J, Pisarska M, Azziz R. How frequently is ultrasound required to diagnose polycystic ovary syndrome in a clinical population? *J Womens Health (Larchmt).* 2024;33:1684-1689. [CrossRef]
16. Pea J, Bryan J, Wan C, et al. Ultrasonographic criteria in the diagnosis of polycystic ovary syndrome: a systematic review and diagnostic meta-analysis. *Hum Reprod Update.* 2024;30:109-130. [CrossRef]
17. Vanden Brink H, Pisch AJ, Lujan ME. A comparison of two- and three-dimensional ultrasonographic methods for evaluation of ovarian follicle counts and classification of polycystic ovarian morphology. *Fertil Steril.* 2021;115:761-770. [CrossRef]
18. van der Ham K, Barbagallo F, van Schilfgaarde E, Lujan ME, Laven JSE, Louwers YV. The additional value of ultrasound markers in the diagnosis of polycystic ovary syndrome. The additional value of ultrasound markers in the diagnosis of polycystic ovary syndrome. *Fertil Steril.* 2025;123:342-349. [CrossRef]
19. Vale-Fernandes E, Pignatelli D, Monteiro MP. Should anti-Müllerian hormone be a diagnosis criterion for polycystic ovary syndrome? An in-depth review of pros and cons. *Eur J Endocrinol.* 2025;192:R29-R43. [CrossRef]
20. Suha SA, Islam MN. A systematic review and future research agenda on detection of polycystic ovary syndrome (PCOS) with computer-aided techniques. *Heliyon.* 2023;9:e20524. [CrossRef]
21. Verma P, Maan P, Gautam R, Arora T. Unveiling the role of artificial intelligence (AI) in polycystic ovary syndrome (PCOS) diagnosis: a comprehensive review. *Reprod Sci.* 2024;31:2901-2915. [CrossRef]
22. Suha SA, Islam MN. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. *Sci Rep.* 2022;12:17123. [CrossRef]
23. Alamoudi A, Khan IU, Aslam N, et al. A deep learning fusion approach to diagnosis the polycystic ovary syndrome (PCOS). *Appl Comput Intell Soft Comput.* 2023;9686697:1-15. [CrossRef]
24. Lv W, Song Y, Fu R, et al. Deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images. *Front Endocrinol (Lausanne).* 2022;12:789878. [CrossRef]
25. Cheng JJ, Mahalingaiah S. Data mining polycystic ovary morphology in electronic medical record ultrasound reports. *Fertil Res Pract.* 2019:5:13. [CrossRef]
26. Nsugbe E. An artificial intelligence-based decision support system for early diagnosis of polycystic ovaries syndrome. *Healthcare Analytics.* 2023;3:100164. [CrossRef]