Artificial intelligence for polycystic ovary syndrome detection and clinical translation

Kaur, G.[a], Sharma, M.[a]

[a]University School of Engineering and Technology, LTSU Punjab, Ropar, India

A B S T R A C T

Polycystic ovary syndrome (PCOS) is difficult to diagnose consistently because clinical, biochemical and ovarian findings vary. Artificial intelligence (AI) can combine these inputs, but high classification accuracy does not necessarily improve diagnosis or patient outcomes. This critical narrative review examines eight purposively selected primary studies published from 2022 to 2025, together with clinical guidance and earlier reviews. We compare prediction targets, data availability, units of analysis, validation independence, reported performance and potential care pathways. Benchmark studies report accuracy approaching 100%, while a hospital electronic health record model reported areas under the curve (AUCs) of 0.80–0.85 across four outcome definitions. More recent studies report external validation for an ultrasound model (AUC 0.967 in 152 patients at another centre) and a laboratory model (AUC 0.923 in 283 patients at another hospital). These findings are promising but cannot be ranked directly because outcomes, sampling strategies and reference standards differ. Polycystic ovarian morphology alone does not establish a PCOS diagnosis. Clinical translation requires calibrated predictions, independent evaluation, appropriate comparison populations, pre specified referral decisions and prospective assessment of benefits and harms. None of the selected studies reported that AI shortened the diagnostic process. Future studies should compare fixed AI models with guideline-concordant usual care and measure patient-centred outcomes rather than accuracy alone. AI and machine learning may support PCOS detection and clinical decision-making, but evidence of clinical benefit is still needed.

Keywords: Polycystic ovary syndrome; Artificial intelligence; Machine learning; External validation; Clinical decision support.

1. Introduction

The prevalence of polycystic ovary syndrome (PCOS) ranges from 8–13% depending on the diagnostic criteria and population (Ghaderzadeh et al., 2025). It has different clinical presentations and some patients are seen more than once for years before a diagnosis is made. Gibson-Helm et al. (2017) found that diagnostic delay and insufficient information were related to dissatisfaction in a survey of women with PCOS. Whether AI can help reduce this delay remains unclear; classifying labelled records or images alone does not establish clinical benefit.

Clinical measurements, ultrasound images and electronic health records (EHRs) have been analysed using artificial intelligence (AI) and machine learning (ML). These systems address different tasks. Classifying an ovarian image, identifying a patient who meets diagnostic criteria and predicting a future recorded diagnosis are different tasks. Performance on one task does not guarantee performance on another. The timing of the input data also matters. Measures taken during a diagnostic work-up might be useful for classification, but offer little evidence of prior identification.

There are several reviews on the existing AI-based models for PCOS. Barrera et al. (2023) included 31 studies but found that only 10 used standardised PCOS diagnostic criteria; Ghaderzadeh et al. (2025) surveyed 80 studies; and Li et al. (2025) catalogued 108 reported features, 12 public datasets and 42 detection tools. It would be redundant to repeat that list of algorithms in another catalogue. Instead, the contribution of the present review is a comparison of evidence to a decision: what does each of these models predict, when does input to each of these models become available, and how independently were each of these models tested, and what additional evidence would be required before each model became evidence to support care.

The review thus differentiates between the classification of ovarian morphology and identification and prediction of the presence of PCOS in cross sectional health records, and the longitudinal health records. It doesn't combine incompatible measures of performance; it does not rank performance

ready numbers. It asks what evidence backs up an acceptable clinical application and which evidence - specifically, earlier diagnosis - has not yet been tested.

2. Review approach

It is a critical narrative review, which is not a systematic review or meta-analysis. This is based on the 2023 international evidence-based guideline for PCOS and the reviews from Barrera et al. (2023), Ghaderzadeh et al. (2025) and Li et al. (2025). Titles, identifiers, sample description and numerical results were validated using a previous source check on 29 September 2026 and publisher full-text pages and PubMed records, supplemented by selected open-access primary articles. Combinations of 'polycystic ovary syndrome' or 'PCOS' with 'machine learning', 'deep learning', 'ultrasound', 'scleral images', 'external validation' and 'multicentre' were used to identify and verify additional examples through a targeted evidence update that took place on 30 October 2026. This was not a count of searches that could be reproduced; there is no claim for a comprehensive search count, PRISMA flow or pooled estimate.

Eight primary studies published between 2022 and 2025 were selected to illustrate different input modalities and levels of validation: Elmannai et al. (2023), Tong et al. (2025), Yao et al. (2025), Suha and Islam (2022), Moral et al. (2024), Zhao et al. (2025), Lv et al. (2022) and Zad et al. (2024). A study was relevant if it was about PCOS classification or identification and the description of a data source, model, and evaluation. Examples were selected to compare benchmark experiments, single centre experiments and independent centre testing but not to provide a representative, or estimate of, the number of examples found in the literature. No main comparison was made for treatment response, fertility prediction and pregnancy outcomes.

For every example, we used the prediction target, data source, number of patients/images (where available), inputs, modelling approach, evaluation environment, and reported metric as well as where applicable, a major clinical drawback. No recalculation of source-reported results was done. The random internal split is described separately, as is the patient level resampling scheme and testing in an independent hospital. ‘Not reported’ indicates missing information, not necessarily that the activity was omitted. No preregistered protocol, duplicate screening, exhaustive assessment of eligibility and independent risk-of-bias scoring or meta-analysis was performed. The organisation of critical questions was done using PROBAST+AI domains; however, it was not used to generate formal study ratings (Moons et al., 2025).

3. AI approaches for PCOS detection

3.1 Clinical and biochemical models

Structured models integrate a patient's menstrual history, anthropometric data and laboratory information. The feasibility is based on the time at which each predictor is available. It is not possible to assume that a tool based on variables collected only after referral can be used to identify individuals at an earlier stage in the care pathway. Ordinary procedures and tests can even vary with different sites, assays and clinical indications (Li et al., 2025).

Elmannai et al. (2023) used a public benchmark of 541 records and 41 attributes. They used a combination of feature selection and stacking ensemble with SHapley Additive exPlanations (SHAP). In its own experiments, the study showed accuracy up to 100 %. This outcome is what the algorithm delivered when the conditions were as they are reported in the benchmark. It does not indicate the performance of the model if it were to be used with a different patient population or measurement procedure or diagnostic reference standard.

Tong et al. (2025) recruited 84 PCOS patients and 75 healthy controls at a hospital. A logistic model, which depended on the hormone level, was developed and obtained an AUC of 0.86 on a validation set of 20 individuals from the same centre. The small validation sample introduces substantial uncertainty. It is also easier to distinguish between established cases and healthy controls than to

distinguish PCOS from patients with other diagnoses worthy of considering, which are generally symptomatic; hence, the finding cannot be interpreted as a result of a population screening.

Yao et al. (2025) assessed 23 routine laboratory indicators in 1,600 women from the First People's Hospital of Jiashan and then externally validated with 283 women from another hospital. They had an internal AUC of 0.919 and external AUC of 0.923 for their XGBoost model; the XGBoost model identified dehydroepiandrosterone sulfate, anti-Müllerian hormone (AMH), triglycerides and age as SHAP important features. This second hospital test was more indicative of transfer than the within-hospital test, but both tests were conducted in the same regional health system. The title of the model does not imply 'early screening', as some of the best predictors come from tests taken during diagnostic investigation. There is no evidence of a decrease in the time to diagnosis (TToD) in the retrospective study (Yao et al., 2025).

3.2 Ultrasound models

Suha and Islam (2022) investigated 594 ovarian ultrasound images (OUSI) using VGG16-based feature extraction and a stacking ensemble with an XGBoost meta-learner, reporting 99.89% accuracy. An image count is not a patient count: if the result is not evaluated independently at the patient level, then it cannot be extrapolated to new women or ultrasound devices.

Moral et al. (2024) developed CystNet using a CNN autoencoder and features extracted from InceptionV3. Their best accuracy using a fully connected classifier was 96.54% and for a random forest, it was 97.75% on the image set they kept for Kaggle. The investigators rejected an overlapping source test partition and repartitioned 1,924 images, and drew out five-fold evaluation and tests on further data. These safeguards and tests are advisory but should not be used in place of prospective assessment of an established diagnostic system in new patients.

Zhao et al. (2025) give a more clinical example of ultrasound. 933 patients provided analysable images (781 in one and 152 in another centre). The system based on YOLOv11 got an AUC of 0.973 in the internal validation sample and 0.967 in the external centre, and the F1 score was 96.9% on both validation samples. The authors outline ovarian-image recognition in the context of a pathway for assessing PCOS, but the external sample is geographically clustered and performance in classifying is not directly related to patient diagnosis, appropriate referral or reduction in delay.

A diagnostic label is especially significant when it comes to imaging studies. A PCOS diagnosis may involve polycystic ovarian morphology (PCOM), but ovarian appearance alone is not diagnostic. In 2023, ultrasound or AMH can be used as supportive procedures to identifying PCOM in adults when indicated, but neither is recommended as a standalone PCOS test, nor are either suggested as a way to define PCOM in adolescents (Teede et al., 2023). Therefore, one should only consider an image-based model to be as a tool to help in the assessment of the ovary if the diagnosis of the patient-level syndrome has been assessed.

3.3 Exploratory non-ovarian imaging

Lv et al. (2022) studied scleral photographs of 388 women with PCOS and 333 women without PCOS in a population of 721 women in China. Eight images per participant were used as a single bag for multi-instance classification which was performed 721 times over participant-level bags using five-fold cross validation. The mean AUC 0.979 and accuracy 92.9% were achieved using a U-Net based segmentation stage and a ResNet based representation. One source of leakage is the tendency for images to be clustered by person. But this method employed a specific acquisition device, an independent hospital validation was not conducted, and an unusual candidate signal was analyzed, whose clinical specificity is unknown. Should be taken as exploratory rather than a modality endorsed as a guideline (Lv et al., 2022).

3.4 Electronic health record models

Zad et al. (2024) analysed data from 30,601 women aged 18–45 at Boston Medical Center, using predictors recorded before each defined outcome. They checked the results of coded and algorithmic derived PCOS outcomes; four definitions, gradient-boosted models yielded AUC values of around

0.85, 0.81, 0.80 and 0.82. Positive and negative predictors reported are associations, not causes. Temporal design increases relevance to an EHR identification use case but there are limitations in the use of coded diagnoses and observed testing patterns as proxies for an independent clinical reference assessment.

While EHR alerts can alert patients to their presence, these alerts are influenced by local documentation and the availability of lab testing data to the algorithm. The absence of a diagnostic code does not establish whether hormone tests were ordered or whether the patient has PCOS. There has to be a pathway of followup and an identified recipient for a deployable alert. No prospective assessment of changes in referrals, diagnostic delay or clinical outcomes were done during the study (Zad et al., 2024).

4. Comparative clinical and translational assessment

Table 1 compares eight purposively selected primary examples. The metrics belong to different targets and sampling schemes and must not be pooled or used to rank modalities. The table also distinguishes a test split from one dataset from evaluation at a second hospital.

Table 1. Representative studies: prediction task, validation and limits of clinical inference

Study / target / sample	Model and key inputs	Evaluation setting	Reported performance	Key translational limitation
Elmannai et al. (2023) / PCOS classification / 541 benchmark records	Optimised feature selection and stacked ML; SHAP	Internal 70:30 and 80:20 benchmark partitions; clinical external test not reported	Best internal accuracy 100%	Public-data performance may depend on cohort, label and preprocessing.
Tong et al. (2025) / case–control PCOS / 84 cases + 75 controls	Hormonal-variable logistic model	20-person same-centre validation subset	AUC 0.86	Healthy controls and small test set do not reflect suspected cases.
Yao et al. (2025) / PCOS case identification / 1,600 development + 283 external patients	Laboratory predictors; XGBoost, SHAP; nomogram	Retrospective; second-hospital external validation	AUC 0.919 internal; 0.923 external	Inputs can be collected after work-up; clinical delay and utility not measured.
Suha & Islam (2022) / ultrasound image classification / 594 images	VGG16 features; stacking ensemble	Image-dataset evaluation; independent patient count unclear	Accuracy 99.89%	Image counts do not show independent patient evaluation.
Moral et al. (2024) / CystNet image classification / 1,924 retained images	InceptionV3, autoencoder; dense or RF classifier	Cross-validation and additional dataset evaluation	Accuracy 96.54% (dense); 97.75% (RF)	Image labels are not a complete PCOS clinical diagnosis.
Zhao et al. (2025) / ovarian ultrasound recognition / 933 patients	YOLOv11 image-analysis model	Prospective, centre 1 n=781; external centre 2 n=152	AUC 0.973 internal; 0.967 external	Improved scan interpretation is not proof of earlier diagnosis.
Lv et al. (2022) / scleral-image PCOS classification / 721 women	U-Net scleral segmentation; ResNet; multi-instance learning	Five-fold participant-bag cross-validation; no new-centre test	Mean AUC 0.979; accuracy 92.9%	Dedicated device and unconventional signal need clinical validation.
Zad et al. (2024) / EHR-based case identification / 30,601 women	Gradient boosting; dated coded and structured record predictors	Single-hospital patient holdout; four label definitions	AUC ~0.85, 0.81, 0.80, 0.82	Local coding and testing practices may bias identification.

AUC, area under the receiver operating characteristic curve; EHR, electronic health record; PCOM, polycystic ovarian morphology; SHAP, SHapley Additive exPlanations. Numbers and performance values are as reported by the cited authors. Independent-centre validation does not in itself establish clinical utility. A reported image count must not be interpreted as an independent patient count.

4.1 Predictive performance and clinical decisions

In their 31 papers, Barrera et al. (2023) reported AUCs of 0.73 – 1.00 and accuracies of 89 – 100%, among which subsets of papers reported AUCs of 0.73 – 1.00 and accuracy of 89 – 100%. These are not the averages of an overall effect. Traditional case–control studies, image-classification experiments and a hospital EHR cohort and independent-centre evaluations are also included in table 1. Their AUCs and accuracies are based on different outcomes, and different case mixes and thresholds, so a modality ranking would be misleading.

Discrimination is a point that shows the ability of a model to distinguish between people with and without the outcome. Calibration: the degree of accuracy of the match between the probabilities predicted and the results obtained. A model can discriminate reasonably well but overestimate the risk. Van Calster et al., (2019) describe the importance of properties in predictions for guiding care. Before a predicted probability is deemed appropriate for clinical interpretation, PCOS studies should include calibration plots, suitable summary measures and uncertainty.

Justification should also be included for the referral threshold. A lower threshold can lead to more potential cases being identified, but may also lead to more unnecessary consultations. Decision-curve analysis can be used to compare the net benefit of various strategies for a range of reasonable thresholds, and take into account the relative consequences of wrongly rejecting and accepting treatments (Vickers and Elkin, 2006). Those strategies must be based on the available care, such as usual assessment for referral, for a PCOS triage model. The decision curve for this was favourable, but needs to be validated in practice.

4.2 External validation and generalisability

The merit of models is variable. Elmannai et al. (2023) conducted internal benchmark experiments, Tong et al. (2025) performed experiments on a small subset of the development hospital, Lv et al. (2022) grouped images into participant-level bags, and Yao et al. (2025) and Zhao et al. (2025) tested in separate institutions. Independent-centre validation is a useful step forward as case mix, sample collection, and equipment can vary, but two centres within one region does not make transportability to all populations/implementation contexts. PROBAST+AI distinguishes between model development and model evaluation, and focuses on participants, predictors, outcomes and analysis (Moons et al., 2025).

It is still necessary to place all repeated scans, multiple eyes, viewing angles, and additional derivatives of a single patient in the same data partition, regardless of the image modality. The scleral study focused explicitly on participant-level image bags (Lv et al., 2022); image studies reporting on the number of files do not provide clarity regarding the test sample. In these cases, the imputation, resampling and feature selection should be performed using the development data within each resampling fold (tabular predictors). Otherwise, information leakage can overestimate results (Moons et al., 2025).

Note that the number of samples should be motivated for the model and the evaluation, and not for a fixed ratio between the number of training and test samples. Riley et al. (2020) explain that development needs are related to the frequency of outcomes, the parameters of the candidates, and the expected performance of the model. The more image features that are derived, the more there are, but not the more independent patients. External studies should also incorporate patients with conditions that are similar to PCOS as this is not the same as comparing patients with entirely different conditions and would only result in diagnostic uncertainty in practice.

4.3 Explainability and clinician interpretation

Explanations need to be assessed. In the Elmannai and Yao studies, saliency maps indicate variables that impact model predictions; image-based systems can, in this context, indicate saliency maps of

image regions. Neither sets the conditions for the predictor to be causal nor does it show that the explanation is stable nor that clinicians can act upon this predictor safely. Ghaderzadeh et al. (2025) showed that explainability methods were only applied in some of the reviewed work. The validity of the explanations should be assessed, as should features which are clinically implausible and the ability of clinicians to detect and correct errors.

In high-stakes situations, Rudin (2019) makes a case for interpretable models if they are performing adequately. An appropriate comparison for PCOS would be a simple clinical model compared with a more complicated algorithm. The evaluation should determine whether the increased complexity has changed the decision or performance sufficiently that the change is worth keeping. Clinicians should be able to review the source observations, and be able to identify missing items and be able to invalidate an inappropriate recommendation. We need to test this empirically to determine the influence of explanations on those actions.

4.4 Data availability and implementation burden

Where all the predictors are available, little additional data collection may be necessary for structured models. If assays are only performed during specialist investigation then that advantage is lost if the routine use is dependent on that. Appropriate imaging models depend on the availability of appropriate imaging data, and the consistent transfer of images into the imaging model (software). EHR models need the ability to extract reliably variables that have dates attached to them and the ability to differentiate between the absence of a finding and the absence of a document. These are practical implications of the input requirements, and not findings on comparative cost-effectiveness.

There is no common basis to compare the costs of implementation in the selected studies. An evaluation of a service might document the time staff need to spend reviewing outputs, performing missing evaluations and dealing with extra referrals. It should also look at who is left out if a model needs full records. DECIDE-AI guides clinicians' early evaluation, including on the role of user in the operating context of a system (Vasey et al., 2022).

4.5 Alignment with diagnostic guidance

The use of a model is restricted by the alignment of guidelines. The 2023 international guideline for the diagnosis of PCOD in adults generally requires 2 of the following: ovulatory dysfunction, clinical/biochemical hyperandrogenism and PCOS following exclusion of mimicking conditions. AMH is not a single diagnostic test, but can be used in the adult diagnostic algorithm instead of ultrasound to define PCOM. If there are irregular cycles, but there is also hyperandrogenism, then ultrasound or AMH are not necessary to make a diagnosis of PCOS. For adolescents, neither ultrasound nor AMH are recommended for diagnosis (Teede et al., 2023). Target label and patient population in an AI study should take these distinctions into consideration.

There is also a need to address measurement and threshold selection in the context of AMH. Bell et al (2022) identified an effect of the AMH substitution on the number of women that were categorised as having PCOS in a community-based sample. A later review by van der Ham et al. (2024) looked at the performance of the different AMH assays, thresholds and populations for diagnosis. Before transferring a model to a new setting, the model should be evaluated using another assay or local threshold. The objectives of the target label should also reflect the diagnostic definition used.

A model for initial triage should apply to information at this stage of caring. If a definitive diagnostic finding is included in the predictors the model is helping in an ongoing assessment. The action to be taken must be clear in the report and the user interface. A flag should not be used for formal diagnosis; however, if symptoms suggest it, a formal diagnosis should not be delayed.

4.6 Evidence-to-decision synthesis

The eight studies are located at various points on an evidence continuum. A public benchmark can provide technical feasibility, a patient independent split can decrease leakage, and independent-hospital testing can uncover some transfer problems. None alone proves that acting on a model output is better than usual care. Other critical next measurements for clinical triage are referrals trig-

gered, number of missed cases, number of false alarms, diagnostic time and patient experience. The objective of a suitable endpoint for ovarian imaging is not the diagnosis of a syndrome, but rather, the quality and efficiency of the morphological assessment by clinicians. The primary interpretive finding of this separation between task and decision is the main one.

There are also some apparent benefits coupled with practical costs. A specialised imaging device is required with the scleral approach, assays are required with the hormone-based systems, there must be indicated scans and compatible acquisition with the ultrasound systems, and there must be consistent records with the EHR systems. A fair comparison should include all the proposed tools in comparison to a plausible baseline using the same patients, reference standard and available information, and reporting missing data and any uncertainty about subgroups. The studies do not provide a common estimate of the cost-effectiveness.

5. Research gaps and future directions

The main research question is not a new point of accuracy on the same familiar dataset, but rather how useful the information is for clinical practice. Future studies of EHR alerts should include patients who have symptoms and/or features recorded in the EHR but have no diagnosis; should include an independent reference assessment; and should prespecify the person who will receive the alert and the action that will follow the alert. An ultrasound assistance study should compare the clinician with and without a fixed model, and report on his diagnosis, as well as reading time, with the same patients. Both designs should not be used as an improved image recognition-only as earlier diagnosis of PCOS.

Independent validation should include clinically realistic comparison groups, including individuals with overlapping reproductive and/or endocrine conditions, and test performance across sites and across clinically relevant demographic and clinical groups. Prespecified thresholds with associated confidence intervals should be presented for sensitivity, specificity, calibration and decision-analytic consequences of these. Often results from external centres like those from Yao et al. (2025) and Zhao et al. (2025) are useful to have as a precursor, but cannot be substituted for prospective testing of clinical benefit, safety and workflow impact. PROBAST+AI can inform design and assessment (Moons et al., 2025).

When multiple modalities add information that enhances the desired decision, it would seem reasonable to combine modalities. There's also additional data-related sources of missingness, opportunities for leakage, and collection expenses. A multimodal model should then be compared with a simpler model with the same patients and reference standard. If either modality is not provided, the system should either have a fall-back mechanism that has been tested or a robust statement that it cannot make a prediction. A comparison like this can demonstrate if having the extra testing was justifiable or not.

Reports should include details of predictor timing, handling of missing data, model selection and evaluation so that others can evaluate the results. TRIPOD+AI offers reporting suggestions for regression and machine learning prediction models (Collins et al., 2024). DECIDE-AI may be used in early deployment studies and CONSORT-AI guidelines (Liu et al., 2020) are recommended in a randomised evaluation of an AI intervention. The standards are used in various phases of research. Checklist does not compensate for poor design or poor choice of diagnostic target.

6. Limitations of this review

This is a purposive narrative review, and the reports selected are not exhaustive and relevant reports may have been overlooked. The targeted October 2026 update was limited to confirming selected accessible papers, and was not a systematic update of all research published during 2026. No retrieval count was available for each database, no formal risk-of-bias grading was available, and no full database search log was available, nor was there any independent screening performed on the databases. Outcome definitions, different diagnostic reference standards, different samples and different

report of the validation of the standards make cross-study inferences difficult. The fact that models contain external validation in two selected studies does not mean they are in routine use and calibrated, safe, equitable or cost-effective. No re-analysis of patient-level data was performed, but results were compared with published reports. The clinical roles proposed are not a framework for evaluation, but rather interpretations.

7. Conclusion

A few of the AI studies that were selected in the context of PCOS show promising retrospective discrimination, and newer external-centre evaluations offer more evidence than individual benchmark results. Ovarian morphology classification does not necessarily replace the diagnosis or identification of syndromes from hospital records, however. Current evidence does not establish that AI shortens diagnostic delays or improves patient outcomes; reported accuracy and cross-site discrimination do not demonstrate these benefits. Future research should define the decision point, incorporate information obtained at that point, test fixed models separately, examine thresholds for calibration and referral, and compare benefit and harm when care is provided in accordance with guidelines versus the fixed models. AI must be used as a tool for making decisions, not a substitute for clinical diagnosis.

References

Barrera, F. J., Brown, E. D. L., Rojo, A., et al. (2023). Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: A systematic review. Frontiers in Endocrinology, 14, Article 1106625. https://doi.org/10.3389/fendo.2023.1106625

Bell, R. J., Islam, R. M., Skiba, M. A., et al. (2022). Substituting serum anti-Müllerian hormone for polycystic ovary morphology increases the number of women diagnosed with polycystic ovary syndrome: A community-based cross-sectional study. Human Reproduction, 37(1), 109–118. https://doi.org/10.1093/humrep/deab232

Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, Article e078378. https://doi.org/10.1136/bmj-2023-078378

Elmannai, H., El-Rashidy, N., Mashal, I., et al. (2023). Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. Diagnostics, 13(8), Article 1506. https://doi.org/10.3390/diagnostics13081506

Ghaderzadeh, M., Garavand, A., & Salehnasab, C. (2025). Artificial intelligence in polycystic ovary syndrome: A systematic review of diagnostic and predictive applications. BMC Medical Informatics and Decision Making, 25, Article 427. https://doi.org/10.1186/s12911-025-03255-6

Gibson-Helm, M., Teede, H., Dunaif, A., & Dokras, A. (2017). Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. The Journal of Clinical Endocrinology & Metabolism, 102(2), 604–612. https://doi.org/10.1210/jc.2016-2963

Li, M., He, Z., Shi, L., et al. (2025). Intelligent detection for polycystic ovary syndrome (PCOS): Taxonomy, datasets and detection tools. Computational and Structural Biotechnology Journal, 27, 1578–1599. https://doi.org/10.1016/j.csbj.2025.04.011

Liu, X., Cruz Rivera, S., Moher, D., et al. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26, 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Lv, W., Song, Y., Fu, R., et al. (2022). Deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images. Frontiers in Endocrinology, 12, Article 789878. https://doi.org/10.3389/fendo.2021.789878

Moons, K. G. M., Damen, J. A. A., Kaul, T., et al. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, Article e082505. https://doi.org/10.1136/bmj-2024-082505

Moral, P., Mustafi, D., Mustafi, A., & Sahana, S. K. (2024). CystNet: An AI-driven model for PCOS detection using multilevel thresholding of ultrasound images. Scientific Reports, 14, Article 25012. https://doi.org/10.1038/s41598-024-75964-3

Riley, R. D., Ensor, J., Snell, K. I. E., et al. (2020). Calculating the sample size required for developing a clinical prediction model. BMJ, 368, Article m441. https://doi.org/10.1136/bmj.m441

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Suha, S. A., & Islam, M. N. (2022). An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Scientific Reports, 12, Article 17123. https://doi.org/10.1038/s41598-022-21724-0

Teede, H. J., Tay, C. T., Laven, J. J. E., et al. (2023). Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. The Journal of Clinical Endocrinology & Metabolism, 108(10), 2447–2469. https://doi.org/10.1210/clinem/dgad463

Tong, C., Wu, Y., Zhuang, Z., & Yu, Y. (2025). A diagnostic model for polycystic ovary syndrome based on machine learning. Scientific Reports, 15, Article 9821. https://doi.org/10.1038/s41598-025-92630-4

van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, Article 230. https://doi.org/10.1186/s12916-019-1466-7

van der Ham, K., Laven, J. S. E., Tay, C. T., et al. (2024). Anti-Müllerian hormone as a diagnostic biomarker for polycystic ovary syndrome and polycystic ovarian morphology: A systematic review and meta-analysis. Fertility and Sterility, 122(4), 727–739. https://doi.org/10.1016/j.fertnstert.2024.05.163

Vasey, B., Nagendran, M., Campbell, B., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28, 924–933. https://doi.org/10.1038/s41591-022-01772-9

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361

Yao, B., Yu, X., Zhang, Y., et al. (2025). Development and validation of an explainable machine learning and nomogram model for early detection and risk stratification of polycystic ovary syndrome: A multicenter study. Frontiers in Endocrinology, 16, Article 1719631. https://doi.org/10.3389/fendo.2025.1719631

Zad, Z., Jiang, V. S., Wolf, A. T., et al. (2024). Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Frontiers in Endocrinology, 15, Article 1298628. https://doi.org/10.3389/fendo.2024.1298628

Zhao, B., Wen, L., Huang, Y., et al. (2025). A deep learning-based automatic recognition model for polycystic ovary ultrasound images. Balkan Medical Journal, 42(5), 419–428. https://doi.org/10.4274/balkanmedj.galenos.2025.2025-5-114