INTERNATIONAL JOURNAL OF
Clinical and Pharmaceutical Innovations
An International Peer Reviewed Open Access Journal

Review Article

ISSN (Online): 3142-8665
CODEN (USA): IJCPSM

2026, Volume 1, Issue 5, Page No: 235-240.

MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE FOR THE DIAGNOSIS OF POLYCYSTIC OVARY SYNDROME (PCOS): A NARRATIVE REVIEW

[1]Abrar Ismail, [2]Dr. Berkheez Shabir, [3]*Dr. Zahoor Hussain Daraz

[1]Department of Computer Applications, Baba Ghulam Shah Badshah University, Rajouri, Jammu & Kashmir, India.
[2]Department of Obstetrics & Gynaecology, Medicare Superspeciality Hospital, Karan Nagar, Srinagar, Jammu & Kashmir, India.
[3]Department of Paediatrics & Neonatology, Medicare Superspeciality Hospital, Karan Nagar, Srinagar, Jammu & Kashmir, India.

Article Info: **Received:** 27 June 2026, **Revised:** 17 July 2026, **Accepted:** 07 August 2026

***Corresponding Author: Dr. Zahoor Hussain Daraz**

Department of Paediatrics & Neonatology, Medicare Superspeciality Hospital, Karan Nagar, Srinagar, Jammu & Kashmir, India.

DOI: https://doi.org/10.5281/zenodo.21893167

Citation:
[1]Abrar Ismail, [2]Dr. Berkheez Shabir, [3]*Dr. Zahoor Hussain Daraz. (2026). Machine Learning And Artificial Intelligence For The Diagnosis Of Polycystic Ovary Syndrome (Pcos): A Narrative Review. International Journal of Clinical and Pharmaceutical Innovations, 1(5), 235-240.

ABSTRACT

Polycystic Ovary Syndrome (PCOS), now broadly termed as Polyendocrine Metabolic Syndrome, is one of the most common endocrine disorders among women of reproductive age, with high global prevalence ranging from approximately 6% to 21% depending on the diagnostic criteria applied, and the majority of cases remaining clinically undiagnosed. Conventional diagnosis relies on the Rotterdam criteria, which require evaluating menstrual irregularity, hyperandrogenism (clinical or biochemical), and polycystic ovarian morphology on USG — a time-consuming, resource-intensive, and subject to inter-observer variability. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) techniques have been widely used to automate and improve the accuracy of PCOS diagnosis. This paper presents a narrative review of recent ML-based approaches for PCOS diagnosis, covering multi-model approaches, deep neural learning architectures, feature selection strategies, and explainable AI techniques. Twelve representative studies published between 2022 and 2025 are examined, with reported diagnostic accuracies ranging from approximately 82.5% to 99.8%. The most frequently used algorithms include Random Forest, Support Vector Machine, XGBoost, and stacked ensemble classifiers, with a Kaggle-hosted dataset of 541 patients from Kerala, India, being the most widely reused public resource. Despite promising results, this review identifies significant gaps, including over-reliance on a small number of public datasets, inconsistent handling of class imbalance, constrained external validation across diverse populations, and limited adoption of standardized Rotterdam-based diagnostic criteria in model design. Recommendations for future perspectives, including region-specific data collection and explainable, clinically deployable models, are discussed.

KEYWORDS: PCOS; Machine Learning; Artificial Intelligence; Random Forest; XGBoost; Infertility; Rotterdam Criteria; USG.

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a challenging and complex endocrine metabolic disorder that affects both adolescents and adult women during their reproductive years. Globally, prevalence is high; however, the estimates vary considerably depending on the diagnostic criteria used, ranging from approximately 6% under the original NIH criteria to over 12% under the broader Rotterdam criteria, with some regional studies reporting figures as high as 21% (Neven et al., 2026; Barrera et al., 2023). According to WHO (World Health Organization), it is estimated that 8–13% of women of reproductive age are affected, and that more than half of these cases go undiagnosed, often for years, due to limited awareness, variable symptom presentation, and the cost of confirmatory testing.

PCOS is clinically significant well beyond fertility concerns. It is strongly associated with anovulatory infertility, insulin resistance, type 2 diabetes, cardiovascular risk, endometrial cancer, and psychological conditions such as anxiety and depression (Neven et al., 2026). Delayed diagnosis — commonly by more than two years — allows these downstream complications to progress before intervention begins (Barrera et al., 2023).

Diagnosis is conventionally guided by the Rotterdam criteria, which require at least two of three features: oligo- or anovulation, clinical or biochemical hyperandrogenism, and polycystic ovarian morphology on transvaginal ultrasound. In practice, this process depends on a combination of menstrual history, hormonal panels (LH, FSH, AMH, testosterone, prolactin), and imaging, interpreted by a specialist — a workflow that is time-consuming and not uniformly accessible, particularly outside urban tertiary care settings.

Machine Learning offers a complementary approach: by learning patterns from clinical, biochemical, and imaging data, ML models can flag high-risk patients earlier, prioritize confirmatory testing, and potentially reduce diagnostic delay. This review surveys recent ML-based approaches to PCOS diagnosis and prediction, summarizes the datasets and algorithms in common use, and identifies the methodological gaps that should inform future research — including work grounded in clinical practice, such as that of gynaecology units managing PCOS-related infertility cases directly.

2. PCOS: Clinical Background and Diagnostic Criteria

Under the Rotterdam consensus, PCOS is diagnosed when at least two of the following three criteria are met, after excluding other causes of hyperandrogenism: (I) oligo-ovulation or anovulation, typically presenting as irregular or absent menstrual cycles; (ii) clinical signs (hirsutism, acne) or biochemical evidence (elevated free testosterone or other androgens) of hyperandrogenism; and (iii) polycystic ovarian morphology on ultrasound, generally defined as 12 or more follicles measuring 2–9 mm in at least one ovary, or increased ovarian volume.

These same clinical, biochemical, and ultrasound-derived variables form the backbone of most ML datasets used in the PCOS literature. Commonly used features include age, menstrual cycle length, body mass index (BMI), hormone levels (TSH, LH, FSH, LH/FSH ratio, AMH, prolactin), signs of hyperandrogenism (hirsutism and acne scores), insulin resistance markers, and follicle counts or ovarian volume from ultrasound. A large EHR-based study additionally found non-gravidity, elevated LH, reduced FSH, obesity, and higher BMI to be strong predictors of PCOS diagnosis in real-world clinical populations (Zad et al., 2024).

3. Review Methodology

This is a conventional literature review rather than a formal PRISMA-based systematic review. Sources were properly identified through targeted searches of peer-reviewed literature (including MDPI, Nature Portfolio, Frontiers, Wiley, and ScienceDirect publications) using combinations of the terms "PCOS," "machine learning," "diagnosis," "prediction," and "classification," with a focus on studies published between 2022 and 2025. Priority was given to peer-reviewed journal articles reporting quantitative diagnostic performance (accuracy, AUC, or equivalent), supplemented by one existing systematic review (Barrera et al., 2023) to provide broader context beyond the studies directly surveyed here. This review does not aim to cover all the fields comprehensively and should be extended with a formal systematic search protocol before being cited as a comprehensive synthesis.

4. Literature Survey

4.1 Ensemble and Optimization-Based Approaches

Panjwani et al. (2025) proposed a hybrid ensemble learning and deep learning architecture (WaOEL) in which the hyperparameters of a seven-classifier ensemble were tuned using a nature-inspired Walrus Optimization algorithm, compared against Cuckoo Search and Random Search variants. Working with a combined symptomatic dataset of 12 attributes, the WaOEL model achieved 92.8% accuracy and an AUC of 0.93, with SHAP-based feature importance analysis identifying obesity and elevated cholesterol as strong indicators of PCOS-positive status.

Mohi Uddin et al. (2025) conducted a comparative analysis of traditional classifiers against ensemble models combined with advanced feature selection techniques, reporting the highest accuracy among all studies surveyed here — 99.78% — using Logistic Regression and SVM. While notable, such near-perfect accuracy on a single dataset warrants caution regarding potential overfitting or data leakage and would benefit from external validation.

An earlier study by Silva et al. (2022), who worked with a smaller clinical cohort of 72 PCOS patients and 73 age-matched controls across 58 clinical and laboratory traits, applied the BorutaShap feature selection technique ahead of a Random Forest classifier, achieving 93% accuracy. The modest sample size limits the generalizability of this result but demonstrates the feasibility of feature-selection-driven approaches on smaller clinical datasets.

4.2 Studies Using the Kaggle Kerala PCOS Dataset

A publicly available dataset of 541 patients collected across ten hospitals in Kerala, India (44 clinical and biochemical features), has become the most frequently reused benchmark in this literature. Multiple studies report strong results on this dataset: a multi-stack ensemble combining Naive Bayes, Decision Trees, KNN, Random Forest, SVM, AdaBoost, Extra Trees, and Gradient Boosting reportedly achieved 98% accuracy with explainable AI methods layered on top, though the authors of subsequent reviews caution that stacking many models on a single dataset raises the risk of overfitting. Other studies using the same dataset report a Linear SVM reaching 91.6% accuracy, and a Multilayer Perceptron combined with an SVM (RBF kernel) reaching 93% accuracy after Pearson-correlation-based feature selection — the latter identifying average follicle size, follicle counts, hirsutism scores, and prolactin levels as the most predictive variables (as summarized in Rahman et al., 2024).

Deep learning has also been applied to this dataset: a study using LSTM, CNN, and a hybrid CNN-LSTM architecture, with SMOTE applied to address class imbalance, reported accuracies of 92.0%, 96.6%, and 94.3% respectively, with corresponding ROC-AUC values in a similar range (as summarized in the Scientific Reports study discussed below).

4.3 Rotterdam-Aligned and Clinical Feature Studies

A 2025 study published in Scientific Reports (Nature Portfolio) took a more clinically structured approach, organizing dataset features into three categories mirroring the Rotterdam criteria: clinical, biochemical, and ultrasound (USG) data. Multiple algorithms — ANN, SVM, Logistic Regression, KNN, and XGBoost — were compared, with XGBoost consistently outperforming the others. This structuring by diagnostic-criteria category is a methodologically stronger design than treating all features as an undifferentiated set, since it better reflects how PCOS is actually diagnosed in clinical practice.

Zad et al. (2024) took a substantially different approach, using electronic health record (EHR) data from 30,601 women aged 18–45 at an urban U.S. safety-net hospital — the largest sample size reported among the studies reviewed here. Four models (Logistic Regression, SVM, Gradient Boosted Trees, and Random Forest) were trained on a combination of ICD-9 diagnostic codes, laboratory values, and radiologic findings. The study found non-gravidity, elevated LH, reduced FSH, obesity, and higher BMI to be the strongest predictors of PCOS diagnosis, and is notable for testing generalizability across a diverse, real-world clinical population rather than a curated research dataset.

4.4 Deployment-Focused and Web-Based Models

Rahman et al. (2024), published in Informatics in Medicine Unlocked, focused specifically on translating a PCOS prediction model into a deployable, web-based tool. Using the Kaggle dataset, a CatBoost classifier achieved 82.5% accuracy on invasive (test-dependent) features and 90.1% on non-invasive clinical indicators alone, while Random Forest and AdaBoost each achieved 94% accuracy. The resulting web interface was designed to accept real-time patient inputs, illustrating a practical path from model development to point-of-care screening.

4.5 Systematic Review Synthesis

Barrera et al. (2023), in the only formal systematic review identified in this survey, screened 135 studies and included 31 in their final synthesis, covering sample sizes ranging from 9 to 2,000 PCOS patients. Support Vector Machines were the most commonly used method (42% of included studies), followed by K-Nearest Neighbours (26%) and regression models (23%). Reported AUC values ranged from 73% to 100% across studies. Critically, the review found that only 32% of included studies used standardized diagnostic criteria such as Rotterdam or NIH, and only 19% completed all three of training, testing, and independent validation — a methodological weakness that this review confirms is still present in several of the more recent studies surveyed above.

5. Comparative Analysis

Table 1 summarizes the datasets, algorithms, and reported performance of the studies discussed in this review.

Table 1: Comparative summary of ML-based PCOS diagnosis studies.

Study (Author, Year)	Dataset / Sample Size	Algorithm(s) Used	Best Reported Accuracy	Key Observation
Panjwani et al., 2025	Combined symptomatic dataset, 12 attributes	Ensemble Learning + Deep Learning, hyperparameters tuned via Walrus Optimization (WaOEL)	92.8% (AUC 0.93)	SHAP-based feature importance; obesity and cholesterol linked to positive prediction
Mohi Uddin et al., 2025	Comparative study, traditional vs. ensemble models	Logistic Regression, SVM, ensemble classifiers with advanced feature selection	99.78%	Highest accuracy among reviewed studies; risk of overfitting not fully addressed
Silva et al., 2022	72 PCOS patients, 73 controls; 58 clinical traits	BorutaShap feature selection + Random Forest	93%	Small sample size limits generalizability
Khanna et al. (as reported in Mohi Uddin et al., 2025; Rahman et al., 2025)	Kaggle Kerala dataset, 541 patients	NB, DT, KNN, RF, SVM, AdaBoost, Extra Trees, Gradient Boost, multi-stack ensemble	98%	Used explainable AI; multi-model stacking on one dataset raises an overfitting concern
Scientific Reports study, 2025 (Nature Portfolio)	Rotterdam-structured dataset (clinical, biochemical, ultrasound features)	ANN, SVM, Logistic Regression, KNN, XGBoost	XGBoost was reported as the best-performing	Feature categories aligned to the Rotterdam criteria for clinical relevance
Abu Adla et al. (as reported in Rahman et al., 2025)	Kaggle PCOS dataset	Linear SVM	91.6%	Simple linear model competitive with complex ensembles
Bharadwaj et al. (as reported in Rahman et al., 2025)	Kaggle Kerala dataset, 541 patients, 44 features	MLP, SVM (RBF kernel), Pearson correlation feature selection	93%	Follicle counts and prolactin were identified as the most predictive features
Ahmad et al. (as reported in Scientific Reports, 2025)	Kaggle PCOS dataset, SMOTE-balanced	LSTM, CNN, hybrid CNN-LSTM	96.6% (CNN)	Deep learning with automated feature engineering; class imbalance explicitly addressed
Rahman et al., 2024 (Informatics in Medicine Unlocked)	Kaggle PCOS dataset	CatBoost, Random Forest, AdaBoost	94% (RF, AdaBoost)	Deployed as a real-time web-based prediction interface
Zad et al., 2024	30,601 patients, hospital EHR data	Logistic Regression, SVM, Gradient Boosted Trees, Random Forest	Highest AUC among EHR-based models	Largest sample size to date; combined ICD-9 codes, labs, and imaging
Barrera et al., 2023 (systematic review)	31 studies pooled, 9–2,000 patients each	SVM (42%), KNN (26%), regression models (23%), others	AUC range 73%–100%	Only 32% of pooled studies used standardized Rotterdam/NIH diagnostic criteria

6. Common Datasets and Features

The single most reused resource in this literature is the Kaggle-hosted PCOS dataset compiled from ten hospitals across Kerala, India, comprising 541 patient records and 44 features spanning clinical, biochemical, and ultrasound variables. At least five of the studies discussed above draw on this same dataset or its derivatives. While its accessibility has accelerated

methodological experimentation, the resulting concentration of published results around a single data source raises concerns about generalizability to other populations, healthcare settings, and ethnic groups — a limitation discussed further in Section 7. Other studies rely on institution-specific clinical cohorts (ranging from under 150 to over 30,000 patients) or on electronic health record extracts, which are less standardized but arguably more representative of real-world diagnostic settings.

7. Research Gaps

Based on the studies reviewed, several recurring limitations can be identified:

1. Dataset concentration: A large share of published models are trained and validated on the same Kaggle Kerala dataset, limiting insight into how these models perform across different populations, ethnicities, and healthcare systems.
2. Inconsistent use of standardized diagnostic criteria: Per Barrera et al. (2023), fewer than a third of studies in the broader literature explicitly ground their target labels in Rotterdam or NIH criteria, which complicates comparison across studies and clinical translation.
3. Small sample sizes in several clinical (non-Kaggle) studies, such as Silva et al. (2022) with 145 total participants, which constrain statistical power and increase the risk of overfitting.
4. Limited external validation: Very few studies test a model trained on one population against an independent, geographically or ethnically distinct cohort. No region-specific study for the Kashmir or wider North Indian population was identified in this review.
5. Uneven adoption of explainability: Only a minority of the reviewed studies (e.g., Panjwani et al., 2025, using SHAP) incorporate explainable AI methods, despite their importance for clinical trust and adoption by practicing gynaecologists.
6. Gap between non-invasive symptomatic screening and full Rotterdam-based diagnosis: Models built purely on low-cost, non-invasive inputs (symptoms, BMI, cycle history) report lower accuracy than those incorporating biochemical or ultrasound data, creating an unresolved trade-off between accessibility and diagnostic rigor.

8. Discussion and Future Directions

The reviewed literature demonstrates that ML methods — particularly ensemble approaches such as Random Forest, XGBoost, and stacked classifiers — can achieve strong diagnostic performance on existing PCOS datasets, in some cases exceeding 95% accuracy. However, the field's heavy reliance on a small number of public datasets, most notably the Kaggle Kerala cohort, means that reported accuracies may not transfer reliably to other clinical populations, including those in Jammu & Kashmir, where no dedicated ML-based PCOS study appears to have been published to date.

Future work would benefit from four directions in particular. First, prospective, multi-center data collection grounded explicitly in Rotterdam criteria — combining clinical, biochemical, and ultrasound features — would allow models to be trained on data that mirrors real diagnostic practice rather than retrofitted symptom sets. Second, external validation across geographically and ethnically distinct cohorts is needed before any model can be considered clinically generalizable. Third, wider adoption of explainable AI techniques (e.g., SHAP, feature-importance analysis) would help build the clinical trust required for adoption by practicing gynaecologists, who must be able to interpret and justify a model's output to patients. Fourth, there is a clear opportunity to translate validated models into lightweight, web- or mobile-based screening tools for use in outpatient and low-resource settings, following the deployment approach demonstrated by Rahman et al. (2024).

A natural extension of this review would be a prospective study conducted in partnership with a gynaecology practice — collecting anonymized clinical, hormonal, and ultrasound data from patients presenting with suspected PCOS or PCOS-related infertility, and using it to train and externally validate a diagnostic support model specific to the regional patient population. Such a study would directly address the dataset concentration and external validation gaps identified above.

9. CONCLUSION

This review has surveyed recent applications of machine learning and artificial intelligence to the diagnosis of Polycystic Ovary Syndrome, examining twelve studies published between 2022 and 2025. Reported diagnostic accuracies ranged from approximately 82.5% to 99.8%, with ensemble methods and gradient-boosted classifiers most consistently achieving strong performance. Despite this progress, the literature remains concentrated around a small number of public datasets, inconsistently grounded in standardized diagnostic criteria, and rarely validated across diverse populations. Addressing these gaps — through prospective, criteria-aligned data collection, external validation, and explainable model design — represents the clearest path toward ML-based tools that can meaningfully support earlier and more consistent PCOS diagnosis in routine clinical practice.

ACKNOWLEDGMENT

The authors acknowledge Medicare Superspeciality Hospital, Karan Nagar, Srinagar, for clinical context and support in framing the diagnostic relevance of this review.

Conflict of Interest

No conflicts of interest.

Funding

This work received no external funding.

REFERENCES

1. Barrera, F. J., Brown, E. D. L., Rojo, A., Obeso, J., Plata, H., Lincango, E. P., Terry, N., Rodríguez-Gutiérrez, R., Hall, J. E., & Shekhar, S. (2023). Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: A systematic review. Frontiers in Endocrinology, 14: 1106625. https://doi.org/10.3389/fendo.2023.1106625
2. Mohi Uddin, K. M., Bhuiyan, M. T. A., Rahman, M. M., Islam, M. M., & Uddin, M. A. (2025). Early PCOS detection: A comparative analysis of traditional and ensemble machine learning models with advanced feature selection. Engineering Reports, 7: e70008. https://doi.org/10.1002/eng2.70008
3. Neven, A. C. H., Forslund, M., Ranasinha, S., Sethi, P., Dhungana, R. R., Mousa, A., Tay, C. T., Teede, H., & Boyle, J. A. (2026). Prevalence of polycystic ovary syndrome: A global and regional systematic review and meta-analysis. Human Reproduction Update, 32(3): 277–312. https://doi.org/10.1093/humupd/dmaf030
4. Panjwani, B., Yadav, J., Mohan, V., Agarwal, N., & Agarwal, S. (2025). Optimized machine learning for the early detection of polycystic ovary syndrome in women. Sensors, 25(4): 1166. https://doi.org/10.3390/s25041166
5. Rahman, M. M., et al. (2024). Empowering early detection: A web-based machine learning approach for PCOS prediction. Informatics in Medicine Unlocked, 47: 101500. https://doi.org/10.1016/j.imu.2024.101500
6. Scientific Reports (Nature Portfolio). (2025). A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features. Scientific Reports. https://doi.org/10.1038/s41598-025-10453-9
7. Silva, I. S., Ferreira, C. N., Costa, L. B. X., et al. (2022). Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. Journal of Endocrinological Investigation, 45: 1–9.
8. World Health Organization. (2021). Breast cancer and related endocrine disorders — WHO fact sheets on reproductive health. https://www.who.int
9. Zad, Z., Jiang, V. S., Wolf, A. T., Wang, T., Cheng, J. J., Paschalidis, I. Ch., & Mahalingaiah, S. (2024). Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Frontiers in Endocrinology, 15: 1298628. https://doi.org/10.3389/fendo.2024.1298628

Note: References attributed "as reported in" a secondary source (e.g., Khanna et al., Abu Adla et al., Bharadwaj et al., Ahmad et al.) were encountered only through their citation in other papers during this review and should be traced to their original publications and verified before this manuscript is submitted for peer review.