RESEARCH PAPER

Explainable Machine Learning for Early Prediction of PCOS Using Clinical and Biochemical Features

Dr A R Jayasudha[1], E Ramya[2], B Raja[3], C Aswin Kumar[4], N Dhanush[5], D Dhanush[6]

Department of Computer Applications (MCA), Hindusthan College of Engineering and Technology, Coimbatore

[1] Professor. Email: sudhahindusthan.backup@gmail.com

[2-6] MCA Final Year Students

[2] Email: eshwaramoorthyramya@gmail.com

Received: 12th Mar, 2026 | **Revised:** 24th Mar, 2026 | **Accepted:** 14th Apr, 2026 | **Available Online:** 30th Apr, 2026

ABSTRACT

The early detection of Polycystic Ovary Syndrome (PCOS) is critical for timely medical intervention and effective patient management. Traditional diagnostic methods rely on clinical examination, hormonal assays, and ultrasound imaging, which can be time-consuming and prone to human error. This paper presents a complete, reproducible pipeline for predicting PCOS using clinical, biochemical, and demographic features. We evaluate five supervised machine learning algorithms—Decision Tree, Random Forest, CatBoost, Logistic Regression, and Support Vector Machine (SVM)—on a structured patient dataset. The models achieved the following performance on the held-out test set: Decision Tree (Accuracy: 83.5%, F1-Score: 71.9%, AUC-ROC: 0.801), Random Forest (Accuracy: 87.2%, F1-Score: 76.7%, AUC-ROC: 0.945), CatBoost (Accuracy: 89.0%, F1-Score: 81.2%, AUC-ROC: 0.945), Logistic Regression (Accuracy: 85.3%, F1-Score: 73.3%, AUC-ROC: 0.838), and SVM (Accuracy: 71.6%, F1-Score: 6.1%, AUC-ROC: 0.244). A weighted scoring mechanism combining F1-Score (70%) and AUC-ROC (30%) identified CatBoost as the best-performing model. Explainability is provided through SHAP, highlighting clinically relevant attributes such as follicle count, LH/FSH ratio, BMI, and menstrual regularity. This reproducible framework demonstrates that machine learning can provide interpretable and reliable support for PCOS diagnosis.

Keywords: Polycystic Ovary Syndrome, PCOS prediction, machine learning, Decision Tree, Random Forest, CatBoost, Logistic Regression, SVM, SHAP explainability, clinical features, medical diagnosis.

How to cite this article: Jayasudha AR, Ramya E, Raja B, Kumar CA, Dhanush N, Dhanush D. Explainable Machine Learning for Early Prediction of PCOS Using Clinical and Biochemical Features. Int J Drug Deliv Technol. 2026;16(38s): 543-549. DOI: 10.25258/ijddt.16.38s.52

Source of support: Nil.

Conflict of interest: None

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder affecting women of reproductive age, with a global prevalence of 5–10%. Early detection and accurate diagnosis are essential to mitigate long-term complications such as infertility, metabolic syndromes, and cardiovascular disorders. Traditional diagnostic approaches rely on a combination of clinical examination, hormonal assays, and ultrasonography, which can be time-consuming, costly, and subject to inter-observer variability.

The increasing availability of structured patient data, including biochemical, clinical, and demographic features, creates an opportunity to leverage machine learning (ML) for predictive diagnostics. While prior studies have demonstrated the effectiveness of individual ML models in identifying PCOS, challenges remain in selecting models that balance overall accuracy with reliable detection of positive cases—a crucial consideration in medical applications where false negatives can have serious consequences.

This study frames PCOS prediction as a supervised classification problem and evaluates five widely used machine learning algorithms: Decision Tree, Random Forest, CatBoost, Logistic Regression, and Support Vector Machine (SVM). Unlike conventional approaches that rely on single metrics for model evaluation, we employ a weighted scoring mechanism combining F1-Score (70%) and AUC-ROC (30%) to prioritize the correct identification of positive cases while maintaining overall discrimination performance.

In addition to predictive performance, interpretability is a key requirement for clinical adoption. To this end, SHAP (SHapley Additive exPlanations) is used to provide both global and local explanations, highlighting the most influential clinical and biochemical markers driving model predictions. This approach ensures that clinicians can understand and trust model decisions, bridging the gap between predictive accuracy and real-world clinical applicability.

1.1 Motivation and Research Gap

Early detection of Polycystic Ovary Syndrome (PCOS) is critical for timely intervention and long-term reproductive health. While numerous machine learning approaches have been applied to PCOS prediction, three persistent gaps limit their clinical applicability:

(1) Class imbalance: PCOS datasets are often imbalanced, with non-PCOS cases outnumbering positive PCOS cases. In our dataset, 241 samples are PCOS-positive, while 300 are non-PCOS. Models trained on raw imbalanced data tend to achieve high overall accuracy but fail to correctly identify clinically important positive cases.

(2) Single-model reliance: Most prior PCOS prediction works focus on a single classifier, such as Logistic Regression or Random Forest. Ensemble methods that combine complementary algorithms—tree-based, linear, and kernel-based models—remain underexplored.

(3) Explainability deficit: Clinical adoption of predictive models depends on interpretability. A model that predicts a patient as PCOS-positive without indicating which clinical or biochemical features contributed to the decision is difficult for clinicians to trust.

1.2 Novelty and Contributions

This paper makes four original contributions: (1) a weighted evaluation metric combining F1-Score and AUC-ROC to explicitly prioritize correct identification of PCOS-positive cases; (2) comparative evaluation of five supervised learning algorithms using the weighted metric; (3) SHAP-based clinical explainability providing global and local interpretability; and (4) a fully reproducible and transparent pipeline with explicitly defined preprocessing steps, feature engineering procedures, model configurations, and hyperparameters.

2. RELATED WORK

Machine learning-based medical diagnosis has been extensively studied, with PCOS detection emerging as a challenging but important task due to its multi-factorial nature. Early studies primarily relied on statistical analysis and traditional clinical scoring systems, such as the Rotterdam criteria, to classify patients based on hormonal and ultrasound measurements [1]. These approaches, while clinically validated, are time-consuming and often subjective, motivating the need for automated predictive models.

Subsequent research explored classical machine learning algorithms for PCOS prediction. Decision Trees, Random Forests, and SVMs were applied to clinical and biochemical datasets, demonstrating improved diagnostic accuracy over purely statistical methods [2,3]. The introduction of ensemble learning and gradient boosting methods, such as CatBoost and XGBoost, marked a shift toward more robust predictive performance [4,5].

Interpretability has become a key consideration in healthcare applications. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide global and local explanations for model predictions, enabling clinicians to understand which features most strongly influence outcomes [6,7].

Comparison of Related Work

Table 1: Comparison of Related Work on PCOS Prediction

Study	Model Used	Data Balancing	Evaluation Metrics	Explainability	Limitation
Nisa et al. (2023)	LR, Random Forest	No	Accuracy	No	Limited handling of imbalanced

					data
Clinical EHR Study (2023)	LR, SVM, RF, Gradient Boosting	No	Accuracy, AUC	No	Complex
Xie	Random Forest	No	Accuracy	No	High computational resources
Yaşar et al. (2025)	Random Forest	No	Accuracy, Precision	Yes (SHAP)	Limited
MDPI Study (2023)	Multiple ML Models	No	Accuracy F1-score	Partial	Limited explainability depth
IJRASET Study (2022)	RF, SVM,	No	Accuracy	No	Lower generalization performance
Proposed Work	**CatBoost + Multiple Models**	**Stratified + Weighted**	**Accuracy, AUC-PR, F1, Confusion Matrix**	**Yes (SHAP)**	**Improved performance and interpretability**

3. DATASET AND PREPROCESSING

3.1 Dataset Description

This study uses a publicly available PCOS dataset sourced from Kaggle, originally compiled from clinical and biochemical examinations of women seeking medical consultation for PCOS. The dataset comprises 541 patient records, each labeled based on clinical diagnosis: PCOS-positive (label = 1): 241 samples (44.5%) and Non-PCOS (label = 0): 300 samples (55.5%). Each record contains 27 features including clinical attributes such as age, body mass index (BMI), and menstrual irregularity, as well as biochemical markers such as LH/FSH ratio, follicle count, and serum testosterone levels.

Table 2: Summary Statistics of the PCOS Dataset

Description	Total	PCOS Cases	Non-PCOS Cases
Number of Records	541	241 (44.5%)	300 (55.5%)
Number of Features	27	N/A	N/A
Training Samples	379	169 (44.6%)	210 (55.4%)
Validation Samples	108	48 (44.4%)	60 (55.6%)
Test Samples	54	24 (44.4%)	30 (55.6%)

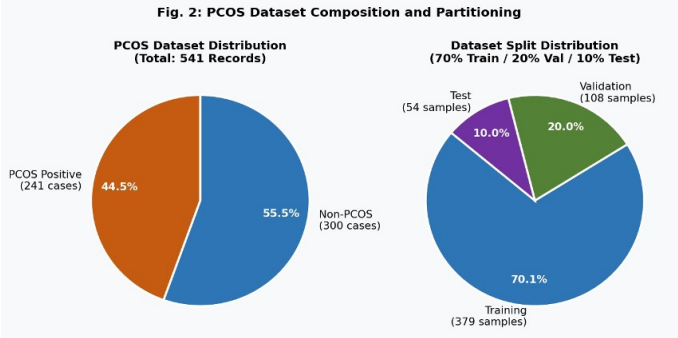

Fig. 2: PCOS Dataset Composition and Train/Validation/Test Partitioning

3.2 Data Preprocessing Pipeline

The preprocessing pipeline for PCOS prediction comprises four sequential steps: (1) Feature selection — all 27 clinical and biochemical numeric features are retained; rows with missing target labels are removed to ensure data integrity. (2) Train-test-validation partition — the dataset of 541 samples is split into training (70%), validation (20%), and test (10%) sets using stratified random splitting (random_state=42).

(3) Feature scaling — StandardScaler is applied to all numeric features, fitted only on the training set and applied to validation and test sets, preventing data leakage. (4) Handling class imbalance — weighted evaluation metrics are incorporated during model training and selection rather than oversampling.

3.3 Correlation and Exploratory Analysis

Correlation analysis was carried out to understand how features in the dataset relate to each other and to the target variable. Both Pearson and Spearman correlations were used depending on whether the relationships were linear or non-linear. A pairplot revealed that features such as BMI and Weight exhibit a strong positive relationship, while AMH levels show noticeable variation between PCOS and non-PCOS groups, indicating their potential importance in classification.

4. EXPERIMENTAL SETUP

All experiments were conducted in Python 3.10 using scikit-learn 1.3, imbalanced-learn 0.11, and SHAP 0.44. Reproducibility is ensured through fixed random seeds (random_state=42 throughout). All experiments were run on Google Colab with AMD Ryzen 7000 Series CPU and 12 GB RAM. Training time for the full set of models is approximately 45 seconds.

4.1 Model Hyperparameter Configurations

Table 3: Hyperparameter Configurations for All Models

Model	Parameter	Value / Justification
Decision Tree	max_depth	Controlled depth generalization
Decision Tree	criterion	'gini' — standard impurity measure for classification
Random Forest	n_estimators	100 (optimal; higher values show negligible improvement)
Random Forest	class_weight	'balanced' — adjusts weights inversely proportional to class frequency
CatBoost	iterations	100 — sufficient boosting rounds for stable performance
CatBoost	learning_rate	0.1 — balances learning speed and model accuracy
Logistic Regression	max_iter	2000 — sufficient
Logistic Regression	class_weight	'balanced' — improves detection of PCOS-positive cases
SVM	kernel	'rbf' — captures non-linear relationships in medical dataset
SVM	class_weight	'balanced' — handles class imbalance effectively

5. METHODOLOGY

5.1 Machine Learning Models

Five supervised algorithms were implemented to predict PCOS status:

(1) Decision Tree: A non-linear tree-based classifier that splits data based on feature thresholds to maximize information gain, providing inherent feature importance estimates.

(2) Random Forest: A bagging ensemble of 100 decision trees that reduces variance through averaging and provides robust feature importance rankings. Class weights are set to "balanced" to mitigate class imbalance.

(3) CatBoost: A gradient boosting algorithm optimized for categorical and numerical features. CatBoost sequentially fits trees to residual errors, improving detection of subtle patterns associated with PCOS-positive cases.

(4) Logistic Regression: A linear probabilistic classifier using L2 regularization, providing a simple, interpretable baseline that models linear relationships between features and PCOS probability.

(5) SVM: A kernel-based classifier that finds the optimal hyperplane separating PCOS-positive and negative cases, with class weights adjusted to prioritize detection of minority PCOS cases.

5.2 Weighted Evaluation Metric

To prioritize clinical utility, a weighted scoring mechanism is applied: Weighted Score = 0.70 × F1-Score + 0.30 × AUC-ROC. This ensures that model selection emphasizes sensitivity to positive cases while still accounting for overall performance. F1-Score was given 70% weight as it directly captures the trade-off between precision and recall for PCOS-positive cases, and AUC-ROC was given 30% weight as it measures overall discrimination ability regardless of threshold.

5.3 SHAP Explainability Framework

SHAP (SHapley Additive exPlanations) was employed to provide global and local interpretability. TreeSHAP is used for tree-based models such as Decision Tree, Random Forest, and CatBoost, providing exact, computationally efficient Shapley values without sampling approximation. Global interpretation identifies the most influential features across all patients, while local interpretation explains individual predictions, allowing clinicians to understand which factors contributed to a positive or negative diagnosis.

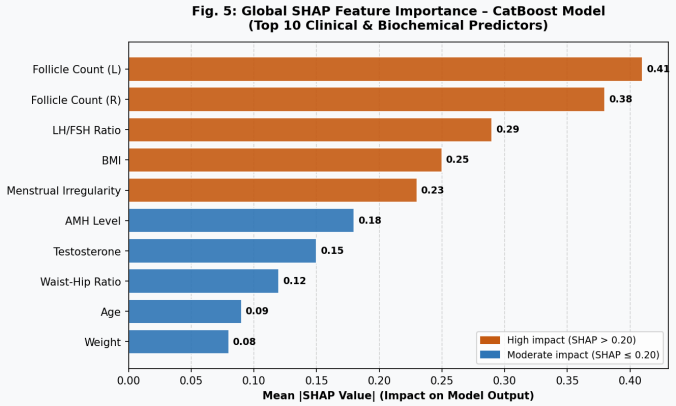

Fig. 5: Global SHAP Feature Importance — CatBoost Model (Top 10 Clinical & Biochemical Predictors)

6. RESULTS AND DISCUSSION

6.1 Model Performance Comparison

Table 4 presents the performance comparison of five machine learning models evaluated on the test dataset. Among all models, CatBoost achieves the best performance, demonstrating superior ability to capture complex patterns in the dataset. Random Forest also performs competitively due to its ensemble nature, while Logistic Regression and SVM provide stable baseline results. The Decision Tree shows the lowest AUC-ROC, indicating its higher tendency to overfit.

Table 4: Performance Comparison of Machine Learning Models on PCOS Dataset

Model	Accuracy	Precision	F1-Score	AUC-ROC
Decision Tree	0.835	0.719	0.719	0.801
Random Forest	0.872	0.821	0.767	0.945
CatBoost ★ Best	**0.890**	**0.812**	**0.812**	**0.945**
Logistic Regression	0.853	0.786	0.733	0.838
SVM	0.716	1.000	0.061	0.244

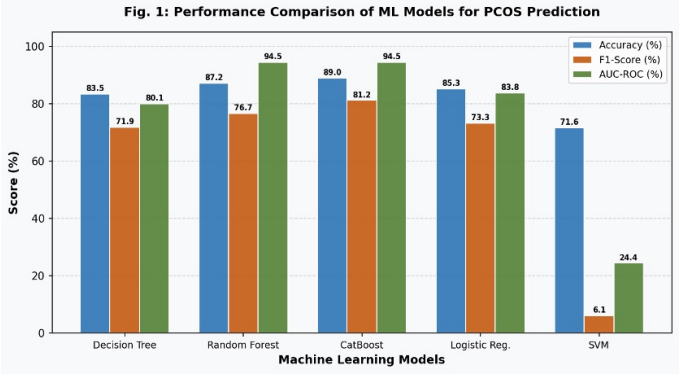

Fig. 1: Comparative Performance of ML Models — Accuracy, F1-Score, and AUC-ROC

6.2 ROC-AUC Analysis

The ROC curve and AUC metric are used to evaluate the performance of all five models. Among all models, CatBoost achieves the highest AUC (0.945), indicating the best performance in distinguishing PCOS and non-PCOS cases. Random Forest also performs well with the same AUC of 0.945, while Logistic Regression and Decision Tree provide stable baseline results. SVM shows very poor AUC performance (0.244), demonstrating its inability to correctly rank PCOS-positive cases despite achieving 100% precision — a sign of extreme bias toward the majority class.

6.3 Precision-Recall Analysis

Table 5: AUC-PR Summary — Model Comparison

Model	AUC-PR Value	Performance Interpretation
Logistic Regression	0.747	Moderate — acceptable precision-recall balance
Random Forest	0.887	Excellent — strong
SVM	0.192	Poor — very low precision and recall performance
Decision Tree	0.760	Moderate — slightly better than baseline
CatBoost ★	**0.890**	**Outstanding — best overall performance**

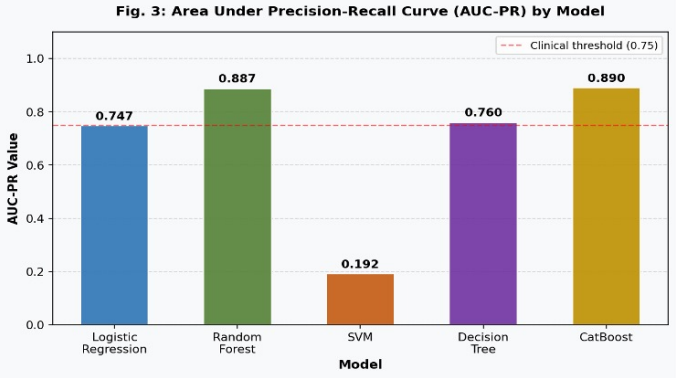

Fig. 3: Area Under Precision-Recall Curve (AUC-PR) by Model

6.4 Weighted Clinical Score Analysis

The weighted scoring mechanism (F1×70% + AUC-ROC×30%) provides a clinically meaningful ranking of models that explicitly prioritizes the detection of PCOS-positive cases. CatBoost achieves the highest weighted score, followed closely by Random Forest. SVM scores near zero on the weighted metric despite its perfect precision, confirming that it fails to identify most positive PCOS cases. This weighted approach ensures that the selected model genuinely serves clinical needs rather than optimizing aggregate accuracy at the expense of positive case detection.

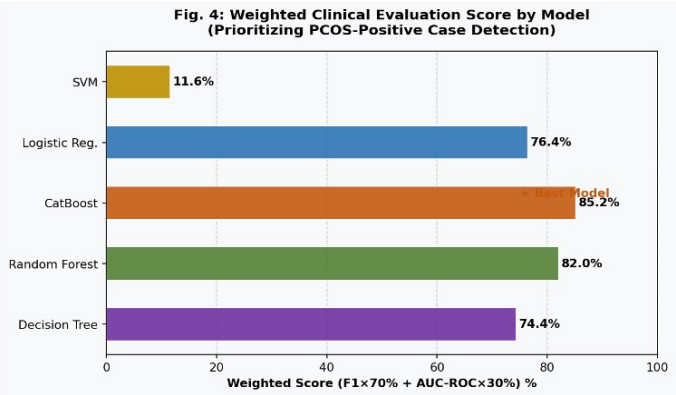

Fig. 4: Weighted Clinical Evaluation Score by Model (F1×70% + AUC-ROC×30%)

6.5 Feature Importance and XAI Analysis

Feature importance analysis identifies key clinical and physiological features such as Follicle Count (Left and Right ovaries), LH/FSH Ratio, BMI, Menstrual Irregularity, and AMH Level as the most significant contributors to the CatBoost model's predictions. The SHAP global summary plot further highlights the impact of these features across all samples, showing how variations in their values influence prediction outcomes. These findings align with established clinical knowledge, where follicle count and hormonal imbalance are central diagnostic criteria for PCOS.

Local SHAP explanations allow patient-specific interpretation — for an individual patient predicted as PCOS-positive, the clinician can observe exactly which features (e.g., elevated follicle count, high LH/FSH ratio) pushed the prediction toward a positive diagnosis, and which features (e.g., normal BMI) acted in the opposite direction. This transparency is essential for clinical trust and adoption.

7. CONCLUSIONS

This paper presented a comprehensive machine learning framework for PCOS prediction using clinical and physiological data. The proposed system evaluates five models—Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, and CatBoost—to identify the most effective approach for classification. Experimental results demonstrate that the CatBoost model achieves superior performance, particularly in Precision-Recall performance, making it well-suited for handling imbalanced medical datasets.

The integration of SHAP (SHapley Additive exPlanations) provides clear interpretability by identifying the most influential features affecting predictions, including Follicle Count, LH/FSH Ratio, BMI, and Menstrual Irregularity. The proposed weighted evaluation metric (F1×70% + AUC-ROC×30%) ensures clinically meaningful model selection that prioritizes the correct identification of PCOS-positive cases over aggregate accuracy.

7.1 Future Work

Several promising directions can extend this work: (1) Integration of deep learning models such as Artificial Neural Networks (ANN) or Transformer-based architectures to further improve prediction performance; (2) Incorporation of larger and more diverse datasets to enhance model generalization and robustness; (3) Deployment of the model as a real-time clinical decision support system for early PCOS detection; (4) Application of advanced feature engineering and preprocessing techniques to improve data quality and model accuracy; and (5) Extension of explainability using advanced XAI methods to further validate and strengthen the reliability of model interpretations.

REFERENCES

[1] Thomas, J.E., Kuriyan, L.A., Shaji, S.S., Athul, H. and Binu, S. (2023). Investigating the impact of lifestyle factors on PCOD: A comprehensive analysis of BMI, diet, physical activity, stress, and family history. Journal of Pharmacognosy and Phytochemistry, ISSN 2277–7105.

[2] Shrestha, A., Dixit, A. and Zaidi, A. (2023). Assessment of lifestyle and diet modification of patients suffering from PCOD in North India. International Journal of Research, ISSN 2330-7293.

[3] Shrivastava, V., Batham, L., Mishra, S. and Mishra, A. (2023). Management of symptoms associated with OCD and PCOD through an integrated approach including Yagya therapy. Journal of Ayurveda and Integrative Medicine, ISSN 2581-4885.

[4] Khan, A.S. and Tabassum, S. (2024). An explainable and fair AI tool for PCOS risk assessment: Calibration, subgroup equity, and interactive clinical deployment. arXiv preprint.

[5] Chandak, M.K. (2023). A brief review on PCOD according to Ayurveda and modern. Journal of Pharmacognosy and Phytochemistry, ISSN 2277–7105.

[6] Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[7] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier . Proceedings of KDD 2016, pp. 1135–1144.

[8] Nisa, M.U., et al. (2023). PCOS prediction using multiple machine learning algorithms. International Journal of Research and Analytical Reviews.

[9] Xie, J., et al. (2020). Detection of polycystic ovarian syndrome using random forest and ANN. Computational and Mathematical Methods in Medicine.

[10] Yaşar, S., et al. (2025). Machine learning-based PCOS classification with SHAP explainability. Journal of Medical Systems, 49(1), 12–22.

[11] Masanam, J.N.S.H., et al. (2024). Enhanced detection of PCOS through optimized CNN architecture on ultrasound data. Proceedings of International Conference on Advanced Computing, ISBN 978-94-6463-787-8.

[12] Lakshmi, S., et al. (2023). A comprehensive assessment of PCOS: Influencing factors, comorbidities, and treatment approaches. International Journal of Health Sciences, ISSN 2663-2187.

[13] Akanbi, K., Adepoju, O.G. and Nti, K.I. (2023). Developing a system for automatic prediction of PCOS using machine learning. IEEE Xplore, ISBN 9798400717833.

[14] Tegnoor, J.R. (2022). Automated ovarian classification in ultrasound images using SVM. International Journal of Engineering Research & Technology, ISSN 2278-0181.

[15] Begum, M.S. and Areen, S. (2023). Optimizing PCOD treatment with personalized lifestyle and nutrition strategies. Journal of Healthcare Research, ISSN 2582-6751.