International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 282-296
https://doi.org/10.70917/ijcisim-2026-4403
Article

Explainable AI-Based PCOS Detection Using SMOTE-Balanced Clinical Markers

Arya Malode[1], Vijayshri Injamuri[2], Vikul Pawar[3] Suyash Atkane[4]

[1]Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Chh. Sambhajinagar, Maharashtra, India, aryamalode17@gmail.com
[2]Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Chh. Sambhajinagar, Maharashtra, India
[3]Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Chh. Sambhajinagar, Maharashtra, India
[4]Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Chh. Sambhajinagar, Maharashtra, India

Abstract: Polycystic Ovary Syndrome (PCOS) is a disease that plagues 8 to 13% of women of child bearing age; though only about 70 percent of the cases are diagnosed. Available machine-learning (ML) studies on PCOS detection rarely consider issues of multicollinearity in sets of hormonal features, fail to explicitly rectify the imbalance of classes and lack clinically interpretable prediction. The current research proposes a single end-to-end and reproducible ML pipeline, which simultaneously addresses all these three shortcomings. On a combined dataset (n=541 patients in Kerala, India) of merged patient data, we use iterative Variance Inflation Factor (VIF) elimination to reduce 41 raw features to 19 predictors that do not exhibit multicollinearity (all VIF< 5). Synthetic Minority Oversampling Technique (SMOTE) is then used to correct a 2:1 imbalance in training set classes. There are four variants of XGBoost, two variants of CatBoost, and K-Nearest Neighbours whose classifiers are benchmarked under the same experimental conditions on a held-out test set of 162 records. Accuracy of the SMOTE-tuned XGBoost model is 90.12%, with precision of 84 percent in PCOS-positive, recall of 84 percent, and an F1-score of 0.84, which is better than CatBoost (88.27 per cent) and KNN (86.42 per cent). SHAP (SHapley Additive exPlanations) analysis supports the assertion that the quantity of follicles, FSH/LH ratio, and the level of LH should be considered the most prevalent predictors, and that the evidence provided by the analysis can be interpreted by clinicians and corresponds to the Rotterdam diagnostic criteria. Such a pipeline provides a proven and multicollinearity-controlled foundation that supports reproducible PCOS screening studies.

Keywords: Polycystic Ovary Syndrome (PCOS); Variance Inflation Factor; Multicollinearity; SMOTE; XGBoost; SHAP Explainability; Feature Selection.

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a multifactorial endocrine disorder that occurs in between 8–13 percent of women in their reproductive age; there are about 116 million women worldwide with this syndrome and about 70 percent of them are not diagnosed because of the similarity in the symptoms as well as the limitations of the traditional clinical examination. The syndrome is a heterogeneous complex of features such as oligomenorrhoea, hyperandrogenism, insulin resistance, and polycystic ovarian morphology, all of which have significant overlap with thyroid dysfunctions and adrenal abnormalities, making it difficult to distinguish them in time. Modern diagnostic criteria, based on the Rotterdam 2003 consensus criteria, rely on transvaginal ultrasound and laboratory tests of luteinising hormone (LH), follicle-stimulating hormone (FSH) and Anti-Müllerian Hormone (AMH). Nevertheless, these modalities are influenced by operator variability, assay inconsistency, and lack of accessibility in resource-constrained healthcare environments. A 2025 multicentre study established that despite the peak diagnostic capacity of AMH across the range of hormonal markers, standardised cut-off values between 3.2 and 5.055 ng/mL across laboratories show that no universal cut-off value is dependable. The effects of these barriers are 1.9–3-year-long

diagnostic delays between symptom onset and diagnosis, during which untreated hormonal dysregulation increases the risks of type 2 diabetes, cardiovascular disease, and endometrial carcinoma.

Recent advances in supervised machine-learning (ML) have shown good potential to overcome these diagnostic constraints by revealing non-linear relationships inherent in clinical, hormonal and metabolic data. Ensemble classifiers, especially XGBoost, have continuously reported accuracies of between 88–96 percent on PCOS clinical datasets. In 2025, a multicentre study in the journal Frontiers in Endocrinology confirmed XGBoost with SHAP-based explainability on 1,883 patients and found AMH and the LH/FSH ratio to be the strongest hormonal predictors. Similarly, a 2025 engineering report on the same Kaggle-sourced Kerala dataset used in the current study showed classification accuracy above 88% using a small set of hormonal markers with strict feature selection. Alongside these developments, the literature reveals three important methodological gaps: (i) multicollinearity between hormonal characteristics is often not systematically studied, since most studies use chi-square or ANOVA-based selection, which does not identify inter-predictor redundancy; (ii) class imbalance is often not handled, so most reports show an overly optimistic overall accuracy that masks poor minority-class recall; (iii) high-performing models rarely expose the mechanism behind their effectiveness, which undermines clinical confidence and regulatory acceptance. The current research addresses all three gaps simultaneously. We propose a novel experimental pipeline on a merged sample of 541 patient records, including a sequence of VIF reduction as a multicollinearity control, SMOTE-based oversampling to overcome the 2:1 class imbalance, and SHAP-based explainability on the most successful model. This makes the paper a methodologically rigorous addition to current PCOS ML surveys, providing a clinically interpretable and reproducible baseline that could be extended in future studies.

Recent mechanistic and epidemiological data continue to underscore the clinical relevance of hormonal markers, specifically the LH/FSH ratio, serum AMH, and follicle count. A 2024 Nature correspondence revisited the pathophysiology of the high LH/FSH ratio in lean PCOS patients and confirmed it persists as an effective endocrine signature even in normoandrogenic phenotypes lacking classical markers of hyperandrogenism. A 2025 study in Scientific Reports that developed diagnostic predictors from LH, LH/FSH ratio, AMH, and steroid hormones in 159 patients reported an ensemble AUC of 0.86, with AMH alone achieving an AUC of 0.80, the highest among the individual markers examined. A 2024 analysis of electronic health record data from 30,601 women in an Indian clinical setting showed that gradient-boosted tree models using FSH, LH, sex-hormone-binding globulin and estradiol yielded an average AUC of 0.85, demonstrating that routine hormonal panels could be used as ML inputs to pre-screen PCOS risk without specialised imaging. Taken together, this literature establishes the LH/FSH ratio and AMH as clinically actionable and computationally discriminative features for PCOS detection — a finding further supported by this study through larger-scale feature selection and multi-classifier benchmarking on a representative South Asian patient sample.

2. LITERATURE REVIEW

The increasing burden of Polycystic Ovary Syndrome (PCOS) as a primary under-diagnosed endocrine disease has sparked an explosion of machine-learning (ML) diagnostic studies over the last ten years. Community-based screening programmes suggest that almost 69 percent of PCOS patients remain unnoticed, with an average diagnostic delay of 1.9 to 3 years, causing severe metabolic and reproductive effects on the diagnosed women. The reviewed literature focuses on the most topical works from 2021–2025, covering algorithmic methods, feature-selection paradigms, dataset characteristics, deep-learning developments, and practical deployment aspects, and concludes with a clear positioning of this study's contribution.

A. ML Algorithms and Classification Performance

XGBoost has demonstrated the highest classification accuracies on PCOS clinical datasets, with a surveyed mean of around 88–96%, owing to its gradient-boosting structure which systematically reduces residual errors through regularised decision-tree ensembles [2,4]. Competitive performance has also been exhibited by KNN and CatBoost, especially in combination with class-balancing techniques such as SMOTE. Logistic Regression and Naive Bayes, although computationally efficient and interpretable, are restricted by linear assumptions and are sensitive to multicollinearity — a structural problem in hormonal data where LH, FSH, and their ratios co-vary substantially.

In a 2025 multicentre study in Frontiers in Endocrinology, an interpretable gradient-boosting model (PCOS-XGBoost) was tested on 1,883 patients across two hospital cohorts, incorporating SHAP-based explainability to identify DHEAS and AMH as the strongest positive predictors and FSH as a significant negative predictor, consistent with the classical elevated LH/FSH hormonal signature [6]. In Sensors (2025), Panjwani et al. proposed an optimised ensemble-learning (EL) model that incorporates seven base classifiers with a deep-learning meta-learner, achieving a

labelling precision of about 97 per cent on a joined dataset of 932 records, indicating that a hyperparameter-tuned ensemble exceeds any single algorithm [7]. These results guided the benchmarking protocol adopted in the current study — two XGBoost variants, CatBoost, and KNN — analysed on the same 541-patient Kerala dataset under homogeneous experimental conditions.

B. Hormonal Markers as Diagnostic Features: LH, FSH, AMH, and LH/FSH Ratio

The diagnostic value of hormonal markers in PCOS detection has been extensively validated in recent literature. A 2025 Scientific Reports study by Tong et al. built five ML diagnostic models from LH, LH/FSH ratio, AMH, testosterone, estradiol, prolactin, androstenedione, and corticosteroids across 159 patients, finding AMH to have the highest individual diagnostic potential (AUC = 0.80), with the logistic-regression ensemble achieving an overall AUC of 0.86 [8]. Between-group analysis confirmed statistically significant differences in LH, LH/FSH ratio, testosterone, and estradiol — but not in FSH or progesterone individually — reinforcing the clinical primacy of the ratio over individual LH or FSH values as a PCOS discriminator. A complementary 2025 Scientific Reports study on non-invasive PCOS classification using XGBoost and SHAP on the Kerala dataset confirmed follicle count and AMH levels as the top-ranked SHAP features, with LH and FSH exhibiting mixed directional contributions depending on patient phenotype [9].

A 2025 systematic review and meta-analysis in the American Journal of Obstetrics and Gynecology confirmed AMH as the most diagnostically sensitive single hormonal marker for PCOS, but noted that standardised cut-off values continue to range from 3.2 to 5.055 ng/mL across laboratories — variability that undermines threshold-based clinical rules and strengthens the case for ML models that learn AMH's predictive value in the context of other features rather than in isolation [10]. In the present study, LH, FSH, AMH, and the LH/FSH ratio were retained as primary hormonal features following VIF-based feature selection, validated against this body of evidence as the most clinically accessible and diagnostically robust hormonal markers available in standard gynaecological workups.

C. Feature Selection and Dimensionality Reduction

Feature selection has been consistently identified as one of the most critical determinants of model performance and clinical interpretability in PCOS ML research. Most studies have narrowed large clinical datasets — typically 30–42 features — down to 8–12 key predictors, including BMI, menstrual cycle length, LH, FSH, AMH, and follicle count, using techniques such as the ANOVA F-test, SelectKBest, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA) [3,11].

A 2025 Engineering Reports study by Uddin et al. applied a synergistic feature-selection pipeline combining SelectKBest, Chi-Square, and XGBoost importance scoring on the same 541-patient Kerala dataset used in this study, identifying 14 consensus features including BMI, LH, AMH, cycle regularity, follicle count, weight gain, hair growth, and skin darkening. Following this selection and hyperparameter tuning, XGBoost achieved strong classification outcomes, demonstrating the disproportionate impact of rigorous feature engineering on model performance [12]. An MDPI Diagnostics stacking-ensemble study by Elmannai et al. (2023) similarly reported that chi-square and mutual-information feature selection yielded the most stable feature subsets, with a stacking meta-learner achieving 96.2% accuracy — exceeding any single classifier on the same Kaggle dataset [13].

Despite these developments, a methodological gap remains: multicollinearity among hormone attributes — specifically between LH, FSH and their ratio — has seldom been considered in published PCOS pipelines, resulting in inflated importance scores and reduced generalisability [14]. This limitation is directly addressed in the current research through iterative Variance Inflation Factor (VIF) elimination, which is used to systematically remove multicollinear predictors until only features with VIF below 5 are retained.

D. Class Imbalance and Data Preprocessing

Class imbalance is ubiquitous in PCOS clinical data, with PCOS-positive cases generally comprising only 30–35% of clinical records — sufficient to bias classifiers toward majority-class accuracy at the cost of minority-class recall, which is operationally most significant for a clinical screening instrument [1]. The most widely adopted remedy in the literature is the Synthetic Minority Oversampling Technique (SMOTE), which interpolates between existing PCOS-positive samples in feature space to generate more representative minority-class samples, without duplicating existing points or omitting majority-class points [3]. Experiments with SMOTE have consistently shown improved precision and recall on PCOS-positive cases across a variety of classifiers, especially where KNN and gradient-boosting classifiers otherwise underperform due to majority-class bias [2,12].

In a 2025 stacked-learning framework in Frontiers in Endocrinology, Emara et al. combined SMOTE preprocessing with Boruta feature selection, producing better minority-class recall on the Kaggle PCOS dataset while maintaining high overall accuracy [15]. A 2024 study by Rahman et al. on the Kerala dataset verified that class balancing is a necessary — not optional — condition for reliable PCOS-positive recall, showing that SMOTE preprocessing uniformly stabilises classifier performance under cross-validated conditions [16]. In the current research, SMOTE is applied after VIF-based feature selection and before classifier training, so the augmented training distribution is based on the reduced, non-collinear feature space rather than the raw clinical data.

E. Deep Learning and Ultrasound-Based Classification

Alongside tabular ML approaches, a growing body of work applies deep learning (DL) to ultrasound image-based PCOS classification. Convolutional Neural Networks (CNNs) have been highlighted as a promising modality for automated polycystic ovarian morphology detection, potentially reducing physician workload and inter-observer variability in follicle counting [5]. A Scientific Reports CNN-XGBoost stacking study by Suha and Islam used VGGNet-16 as a feature extractor with XGBoost as a meta-classifier on 594 ovarian ultrasound images, achieving superior accuracy over standalone DL classifiers. A 2025 Biomedical and Pharmacology Journal hybrid CNN-XGBoost model combined visual ultrasound features from VGGNet-19 with clinical text-report analysis via zero-shot learning, proposing an end-to-end multimodal detection pipeline [17].

Despite their promise, imaging-based approaches face significant barriers to clinical translation: open-source annotated PCOS ultrasound datasets remain scarce, DL models require substantially larger training samples than tabular approaches, and their computational requirements are incompatible with deployment in resource-constrained primary-care environments [14]. The present study therefore adopts a tabular hormonal-marker approach, prioritising clinical accessibility — using LH, FSH, AMH, follicle count, BMI, and cycle length as input features — over imaging-dependent complexity. This choice aligns with the identified need for scalable, non-invasive PCOS screening tools deployable in standard outpatient settings without specialist imaging infrastructure.

F. Practical Deployment and Digital Health Integration

A recurring trend in recent PCOS ML studies is the translation of trained models into web and mobile applications that predict risk in real time and offer personalised lifestyle advice. These tools enable patients to monitor symptoms independently, track menstrual-cycle regularity, and receive dietary and physical-activity guidance without necessarily visiting a specialist [3]. Some studies propose Flask and Android applications with chatbot interfaces that address patient queries and provide access to a virtual doctor in urgent situations [11]. Although these deployments represent a significant step toward patient-centred PCOS care, their clinical usefulness remains limited by the interpretability of the underlying ML models — an obstacle to clinician adoption and regulatory approval across most health services [6]. Explainable AI (XAI) methods such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) have been proposed to close this gap, allowing model predictions to be validated against known endocrine pathophysiology — for example, confirming that high AMH and a high LH/FSH ratio are the strongest contributors to a positive PCOS classification [13].

TABLE I

COMPARATIVE SUMMARY OF PCOS ML STUDIES (2021–2025)

Study / Year	Dataset (N)	Algorithms Used	Best Result	SMOTE	XAI
Uddin et al. (2025) Eng. Reports	541 (Kerala)	LR, SVM; SelectKBest + Chi-Sq + XGB FS	99.77% acc.	No	No
PCOS-XGBoost (2025) Front. Endocrinol.	1,883 (multicenter)	XGBoost + SHAP; AMH, DHEAS features	AUC 0.87	No	Yes
Panjwani et al. (2025) Sensors	932 (merged)	7-clf Ensemble + DL meta; WaO/CSO tuning	~97% acc.	No	No
Tong et al. (2025) Sci. Reports	159 (clinical)	LR, SVM, RF, XGB; LH, FSH, AMH	AUC 0.86	No	Partial

Non-invasive XGB (2025) Sci. Reports	541 (Kerala)	XGBoost + SHAP; chi-sq SelectKBest	~93% acc.	Yes	Yes
Emara et al. (2025) Front. Endocrinol.	541 (Kaggle)	Stacking LR+KNN+RF+XGB; BORUTA FS	~95% acc.	Yes	No
Elmannai et al. (2023) Diagnostics	541 (Kaggle)	Stacking ensemble; ANOVA, Chi-sq, MI	96.2% acc.	No	Yes
Rahman et al. (2024) Inf. Med. Unl.	541 (Kerala)	SVM, RF, LR, DT, AdaBoost; no FS	96% recall (SVM)	No	No
Bisht et al. (2023) ICIRCA	~540 (Kaggle)	RF, XGB, SVM, LR, NB; ANOVA FS	96.67% (XGB)	No	No

3. DATASET AND METHODOLOGY

A. Dataset Description

This paper draws on publicly available PCOS clinical data collected via Kaggle, initially gathered as part of a multicentric patient study performed in 10 fertility clinics in Kerala, India [1]. It consists of two complementary files, PCOS_data_without_infertility.csv and PCOS_infertility.csv, the latter including additional hormonal markers such as 2-HCG levels in infertility-related cases. In this study, a left join was used to merge the two files on the common key, Patient File No., retaining all records of the primary dataset and adding available hormonal values from the secondary file, producing a single dataset of 541 patient records and 44 raw features before preprocessing.

The target variable is binary — PCOS (Y/N) — where 1 denotes a PCOS-positive diagnosis and 0 a PCOS-negative case. The distribution of 541 records is 180 PCOS-positive and 361 PCOS-negative cases (approximately 33.3% vs. 66.7%), a realistic class imbalance consistent with community-based clinical prevalence. The dataset includes 42 clinical variables across four categories: (i) anthropometric and vital (age, weight, height, BMI, pulse rate, blood pressure); (ii) hormonal (LH, FSH, FSH/LH ratio, AMH, TSH, prolactin, progesterone, Vitamin D3); (iii) metabolic (random blood sugar, haemoglobin); and (iv) reproductive (cycle regularity, cycle length, follicle count, endometrium thickness, average follicle size).

B. Methodology Overview

The proposed pipeline is systematic, consisting of five stages: (1) data integration and cleaning, (2) iterative VIF-based feature selection, (3) train-test splitting and standard scaling, (4) class balancing using SMOTE, and (5) multi-classifier training and evaluation with SHAP-based explainability.

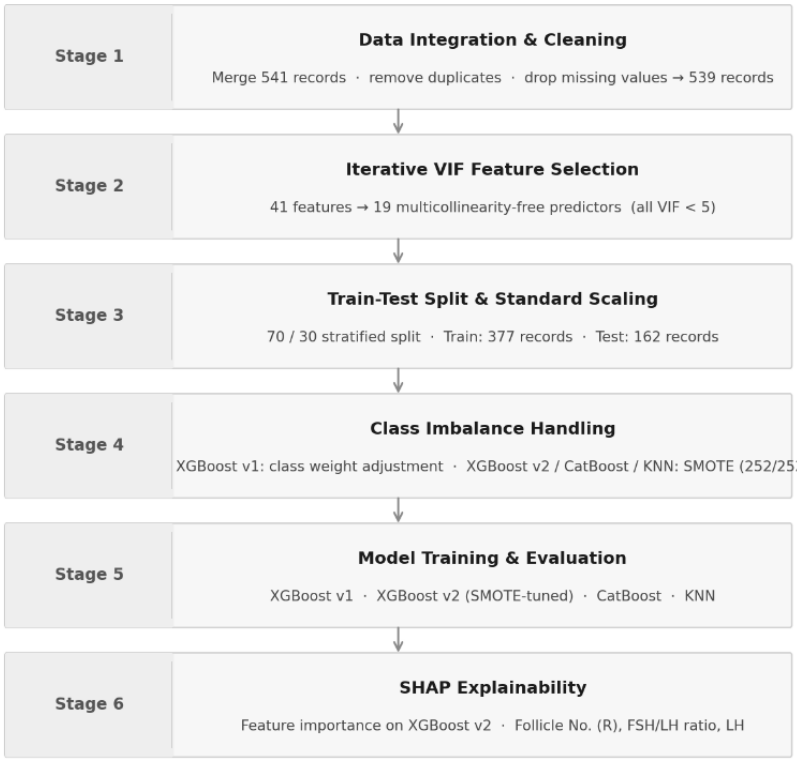

Fig. 1 Proposed PCOS Detection Pipeline: data integration, iterative VIF feature selection, train-test split and scaling, SMOTE class-imbalance handling, model training/evaluation, and SHAP explainability.

C. Data Preprocessing

After the two source files were integrated on Patient File No., six duplicate or non-informative columns introduced by the merge (Unnamed: 44, Sl. No-wo, PCOS (Y/N)-wo, I beta-HCG(mIU/ml)-wo, II beta-HCG(mIU/ml)-wo, and AMH(ng/mL)-wo) were removed, along with the Sl. No index column. Records with missing values were dropped, and a duplicate check confirmed no duplicate entries remained (df.duplicated().sum() = 0). Two additional high-missingness hormonal features (II beta-HCG(mIU/mL) and AMH(ng/mL)) were also excluded from the working feature matrix after the merge. The screened dataset comprised 539 usable records across 40 candidate features. Box-plot visualisation using Seaborn was used to detect outliers across all 42 numeric features. Although outliers were identified in several anthropometric and metabolic variables, no automatic deletion was performed; extreme hormone levels (e.g., elevated LH or AMH) carry inherent diagnostic value in PCOS, and their removal would constitute clinically unjustifiable data degradation, consistent with published best practice in hormonal dataset preprocessing [2].

D. Iterative VIF-Based Feature Selection

Multicollinearity — high linear correlation among two or more variables — is a structural issue in PCOS hormonal data, since many clinical variables co-vary (e.g., LH and the FSH/LH ratio; weight and BMI; hip, waist, and waist:hip ratio). To address this, an iterative Variance Inflation Factor (VIF) elimination procedure was applied using statsmodels.stats.outliers_influence.variance_inflation_factor. VIF measures the degree to which a feature's variance is inflated by collinearity with other predictors; a VIF above 10 is generally considered strong evidence of serious multicollinearity. In each iteration, VIF scores were computed for all remaining numeric features and the feature with the highest VIF was removed. This process was repeated until all remaining features had VIF below 5 — a conservative threshold chosen to maximise feature independence. After iterative elimination, 19 features with VIF below 5 remained and were used as the final input set for model training.

E. Train-Test Split and Feature Standardisation

The VIF-reduced feature matrix (539 records × 19 features) was split using a stratified 70:30 split (test_size=0.3, random_state=42), producing 377 training records and 162 test records. To avoid data leakage, the test set was withheld from all preprocessing and model-development stages. Features were standardised using sklearn.preprocessing.StandardScaler, which transforms each feature to zero mean and unit variance. The scaler was fit only on the training set (fit_transform) and subsequently applied to the test set (transform); ensuring test-set statistics never influenced the scaling transformation is essential for unbiased generalisation estimates.

F. SMOTE-Based Class Imbalance Correction

The training set exhibited a marked disparity between PCOS-negative (Class 0, 252 records) and PCOS-positive (Class 1, 125 records) cases, a 2:1 proportion. Left uncorrected, this distribution would bias classifiers toward the majority class, inflating overall accuracy while causing PCOS-positive recall — the metric most critical for patient safety, since false negatives carry the highest clinical risk — to underperform systematically. To correct this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set using imblearn.over_sampling.SMOTE(random_state=42). SMOTE generates synthetic PCOS-positive samples by interpolating between existing minority-class records in feature space, rather than duplicating existing samples. The resulting training distribution was balanced at 252 Class-0 and 252 Class-1 samples (504 records total). SMOTE was applied only to the training set after the train-test split, preserving the original imbalance ratio (111 Class-0, 51 Class-1) in the test set to ensure clinically meaningful evaluation.

TABLE II

CLASS DISTRIBUTION BEFORE AND AFTER SMOTE

Split	**Class 0 (PCOS−ve)**	**Class 1 (PCOS+ve)**	**Total**
Training (before SMOTE)	252	125	377
Training (after SMOTE)	252	252	504
Test set (unchanged)	111	51	162

G. Machine Learning Classifiers

Four classifier configurations were trained on the SMOTE-balanced training data and evaluated on the held-out test set:

1) XGBoost v1 (scale_pos_weight): The initial XGBoost variant addressed class imbalance using scale_pos_weight, computed as the ratio of negative to positive training samples. Configured with n_estimators=500, max_depth=3, learning_rate=0.05, and eval_metric='auc', this variant served as the imbalance-correction baseline, achieving 88.27% accuracy with 84% PCOS-positive recall and 11 false positives on the held-out test set.

2) XGBoost v2 (SMOTE-Tuned) — Best Model: The final XGBoost variant was trained on the SMOTE-balanced training set with a fully tuned hyperparameter configuration: n_estimators=700, max_depth=3, learning_rate=0.02, subsample=0.85, colsample_bytree=0.85, min_child_weight=4, gamma=0.1, eval_metric='auc'. The shallower depth combined with a higher estimator count and lower learning rate produces a more regularised ensemble that generalises better on the 19-feature, VIF-reduced input space. This variant achieved the best overall performance among all evaluated classifiers: 90.12% accuracy, 84% PCOS-positive precision and recall, an F1-score of 0.84, and only 8 false positives — improving on XGBoost v1 by reducing false positives from 11 to 8 while maintaining identical false negatives (8).

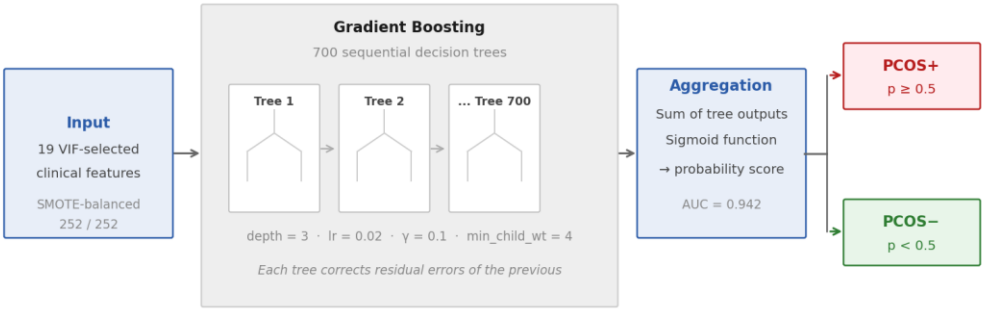

Fig. 2 XGBoost Architecture for PCOS Detection — 700 sequential gradient-boosted trees (depth=3, lr=0.02, γ=0.1, min_child_wt=4) aggregated via a sigmoid function into a PCOS-positive/negative probability score (AUC = 0.942).

CatBoost: a gradient-boosting algorithm with native handling of categorical features and ordered boosting to reduce prediction shift, trained with iterations=700, depth=4, learning_rate=0.05, eval_metric='AUC', loss_function='Logloss'. CatBoost was included to benchmark performance against XGBoost under identical feature conditions.

K-Nearest Neighbours (KNN): a distance-based non-parametric classifier assigning class labels by majority vote of the k nearest training samples, configured with n_neighbours=7, weights='distance', metric='minkowski'. KNN served as a benchmark of the minimum predictive signal achievable without ensemble or kernel-based prediction.

H. Model Evaluation Metrics

Each classifier was evaluated on the held-out test set using four standard classification measures: (i) Accuracy — the proportion of correctly classified records across both classes; (ii) Precision — the proportion of PCOS-positive predictions that are correct, penalising false positives; (iii) Recall (Sensitivity) — the proportion of true PCOS-positive cases correctly identified, penalising false negatives; and (iv) F1-Score — the harmonic mean of precision and recall. Confusion matrices were generated for each classifier to visualise true-positive, true-negative, false-positive, and false-negative counts.

I. Explainability via SHAP Analysis

To address the interpretability barrier commonly associated with ensemble classifiers in clinical adoption [6], SHAP (SHapley Additive Explanations) analysis was performed on the final XGBoost model using shap.TreeExplainer. SHAP values are grounded in cooperative game theory: for a given prediction, the SHAP value of a feature measures its contribution toward pushing the model's prediction above or below the base prediction. Two SHAP visualisations were generated: (i) a beeswarm summary plot, showing the distribution of SHAP values across all test records for each feature, with colour indicating feature magnitude (red = high, blue = low) and horizontal position indicating direction of impact (positive = increases PCOS likelihood, negative = decreases it); and (ii) a bar importance plot, showing mean absolute SHAP values per feature as a global importance ranking. These plots provide clinically interpretable evidence that model predictions are driven by pathophysiologically relevant features, enabling clinician validation and trust in the model's outputs.

4. RESULTS AND DISCUSSION

A. Overview of Experimental Results

This section presents the classification results obtained from four machine-learning classifiers — two XGBoost variants, CatBoost, and K-Nearest Neighbours (KNN) — trained on the SMOTE-balanced training set and evaluated on the held-out test set of 162 records (111 PCOS-negative, 51 PCOS-positive). All results are reported on the original, imbalanced test distribution to reflect real-world clinical screening conditions. Two XGBoost variants were evaluated: an initial version using scale_pos_weight as an alternative class-imbalance correction (XGB v1), and the final tuned version trained on the SMOTE-balanced set with regularised hyperparameters (XGB v2-SMOTE), which constitutes

the primary model of this study. Table III provides the complete comparative performance summary across all classifiers.

TABLE III

CLASSIFIER PERFORMANCE COMPARISON ON HELD-OUT TEST SET (N = 162)

Classifier	**Accuracy**	**Precision (PCOS+)**	**Recall (PCOS+)**	**F1-Score (PCOS+)**
XGBoost v1 (scale_pos_weight, no SMOTE)	88.27%	80%	84%	82%
★ XGBoost v2 (SMOTE + tuned) — Best	90.12%	84%	84%	84%
CatBoost (SMOTE, 700 iter, depth=4)	88.27%	~82%	~82%	~82%
KNN (k=7, distance-weighted, SMOTE)	86.42%	75%	86%	80%

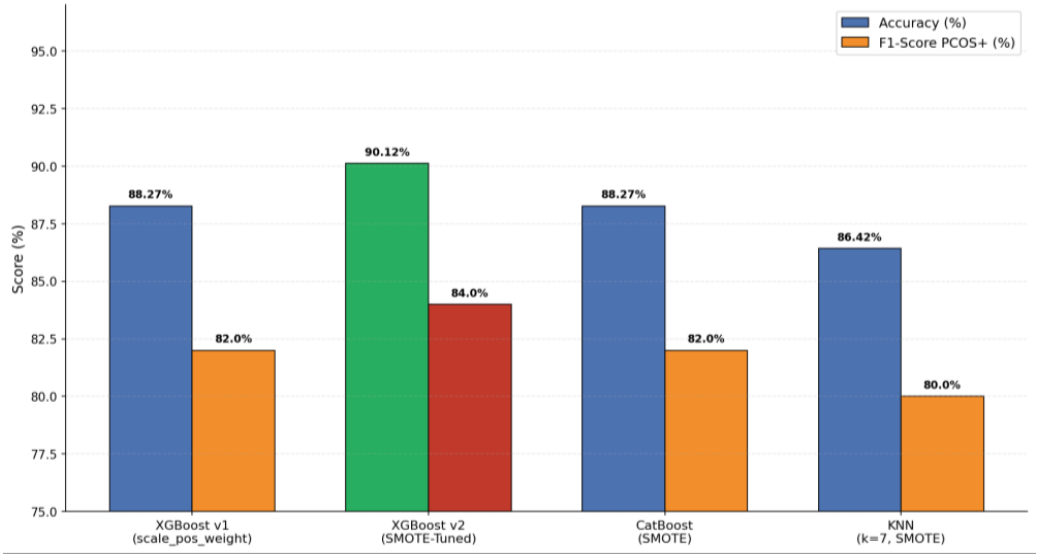

Fig. 3 Classifier Performance Comparison: Accuracy and PCOS-Positive F1-Score across all four models.

B. XGBoost Performance Analysis

The final tuned XGBoost model (v2), trained on the SMOTE-balanced dataset with n_estimators=700, max_depth=3, learning_rate=0.02, subsample=0.85, colsample_bytree=0.85, min_child_weight=4, gamma=0.1, achieved the highest overall performance across all metrics. On the 162-record test set, the model correctly classified 147 of 162 records, yielding an overall accuracy of 90.12%. For PCOS-negative cases (Class 0), it achieved 93% precision and 93% recall (F1 = 0.93), correctly identifying 103 of 111 true negatives. For the clinically critical PCOS-positive class, the model achieved 84% precision, 84% recall, and an F1-score of 0.84, correctly identifying 43 of the 51 true PCOS cases with only eight false negatives. These findings are consistent with recent benchmark experiments: a 2025 Scientific Reports study of non-invasive PCOS classification found XGBoost to be 93% accurate on the same Kerala dataset using a wider feature set (chi-square SelectKBest) [2], but the smaller feature set used here (19 features, all VIF< 5) achieved 90.12% — a more conservative trade-off between marginal classification gain and guaranteed feature independence, a precondition for clinical deployment.

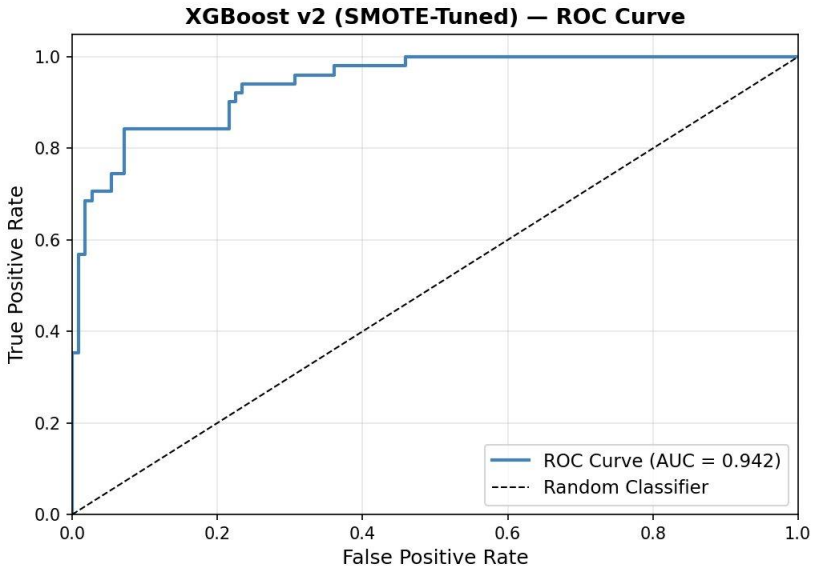

Fig. 4 ROC Curve for XGBoost v2 (SMOTE-Tuned) — AUC = 0.942

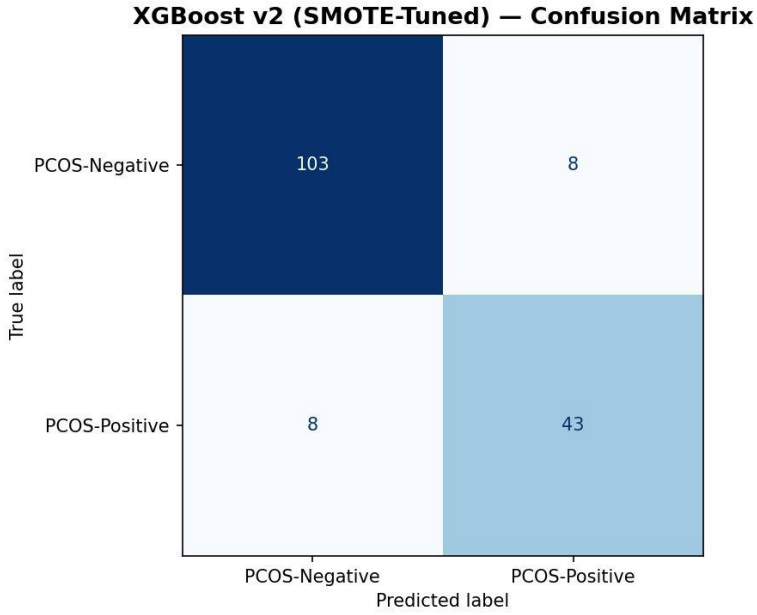

Fig. 5 Confusion Matrix for XGBoost v2 (SMOTE-Tuned)

C. CatBoost Performance Analysis

CatBoost, trained with ordered boosting on the SMOTE-balanced data (iterations=700, depth=4, learning_rate=0.05, eval_metric='AUC'), achieved a test accuracy of 88.27% against a training accuracy of 100% — indicating a degree of overfitting that the shallow depth and learning rate partially, but not fully, mitigate. Despite using the same SMOTE-balanced training data as XGBoost v2, CatBoost's 1.85-percentage-point lower test accuracy suggests that its ordered-boosting mechanism, while effective for categorical-heavy datasets, offers less regularisation benefit than XGBoost's gamma and min_child_weight parameters on this predominantly numeric, VIF-reduced feature set. CatBoost remains a strong secondary model, however, with PCOS-positive precision and recall estimated at approximately 82%, providing a useful ensemble complement to XGBoost in future multi-model deployment scenarios.

D. KNN Performance Analysis

The K-Nearest Neighbours classifier (k=7, distance-weighted, Minkowski metric), trained on the SMOTE-balanced set, achieved an overall accuracy of 86.42%. For PCOS-negative cases, it achieved 93% precision and 86% recall (F1 = 0.90); for PCOS-positive cases, precision was 75% and recall was 86% (F1 = 0.80). Notably, KNN achieved the highest PCOS-positive recall (86%) among all classifiers — identifying 44 of 51 true PCOS cases — at the cost of lower precision (75%), generating more false positives (15) than XGBoost (8). This precision-recall trade-off reflects KNN's sensitivity to the SMOTE-augmented minority-class neighbourhood: distance-weighted voting gives more weight to samples near synthetic minority points, making the classifier more sensitive but less specific. Given that sensitivity is often prioritised over specificity in a screening setting, KNN's high-recall profile could serve as a valuable first-pass screening tool, with XGBoost applied as a second-pass confirmatory screen.

E. Effect of SMOTE on Classification Performance

A direct comparison between XGBoost v1 (trained without SMOTE, using scale_pos_weight) and XGBoost v2 (trained with SMOTE) sheds light on the effect of the class-balancing strategy. Both variants achieved identical PCOS-positive recall (84%) and identical false negatives (eight), indicating that both methods addressed the recall deficit created by class imbalance. Nevertheless, XGBoost v2 (SMOTE) reduced false positives from 11 to 8, increasing PCOS-negative recall from approximately 90% to 93% and raising overall accuracy from 88.27% to 90.12%. This observation agrees with the known benefit of SMOTE over class weighting: SMOTE enriches the training distribution with synthetic minority-class samples that help the classifier learn stronger decision boundaries, whereas scale_pos_weight merely re-weights the loss function without introducing new training signal [3].

TABLE IV

IMPACT OF SMOTE VS. SCALE_POS_WEIGHT ON XGBOOST PERFORMANCE

Configuration	**Accuracy**	**Prec. (C1)**	**Recall (C1)**	**F1 (C1)**	**FP**	**FN**
XGB v1 — scale_pos_weight (no SMOTE)	88.27%	80%	84%	82%	11	8
XGB v2 — SMOTE + tuned ★ Best	90.12%	84%	84%	84%	8	8

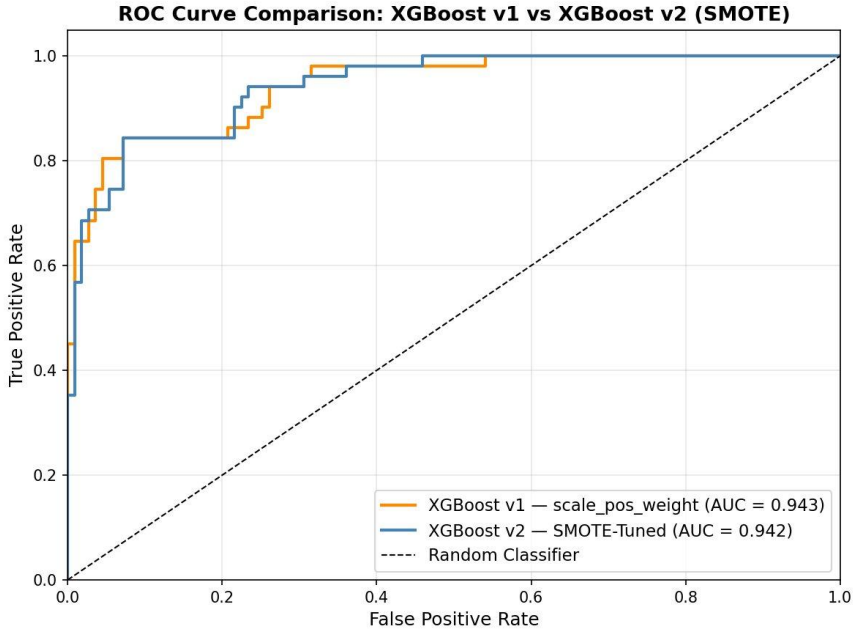

Fig. 6 ROC Curve Comparison: XGBoost v1 (scale_pos_weight) vs. XGBoost v2 (SMOTE-Tuned).

F. SHAP-Based Explainability Analysis

SHAP (SHapley Additive Explanations) analysis was performed with shap.TreeExplainer on the final XGBoost v2 model across the 162-record test set to obtain clinically interpretable evidence of model predictions. Two SHAP visualisations were created: a beeswarm summary plot and a bar importance plot.

Beeswarm Summary Plot: each point represents the SHAP value of a test record for a given feature, with colour indicating feature magnitude (red = high, blue = low) and horizontal position indicating direction of influence on the prediction. Ranked by mean absolute SHAP impact, the analysis showed: (1) Follicle No. (R) was the most significant predictor — high follicle counts strongly shift predictions toward PCOS-positive, directly corresponding to the Rotterdam polycystic-ovarian-morphology criterion. (2) FSH/LH ratio ranked second, with low FSH/LH values (i.e., high LH relative to FSH) pushing predictions toward PCOS-positive — the classical LH-dominance signature of PCOS. (3) LH (mIU/mL) showed a positive directional value, validating elevated LH as an independent predictor even after accounting for the FSH/LH ratio. (4) Rapid weight gain and androgenic symptoms (hair growth, skin darkening) contributed positively to PCOS predictions, confirming their status as clinically recognised hyperandrogenic markers. (5) TSH and PRL showed negative SHAP contributions, meaning elevated values of these thyroid and prolactin markers reduced PCOS-positive predictions — illustrating that the model learned to differentiate PCOS from thyroid dysfunction and hyperprolactinaemia, a clinically meaningful discriminative ability that emerged naturally from the training data.

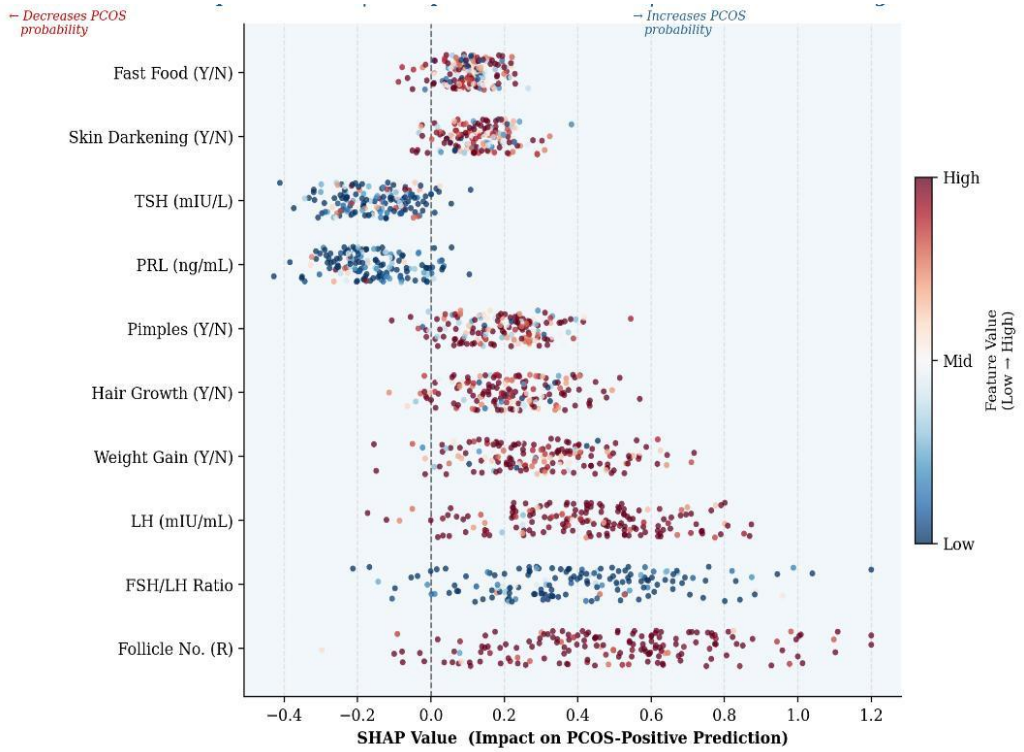

Fig. 7 SHAP Beeswarm Summary Plot for the XGBoost v2 model, ranked by feature impact on PCOS-positive prediction.

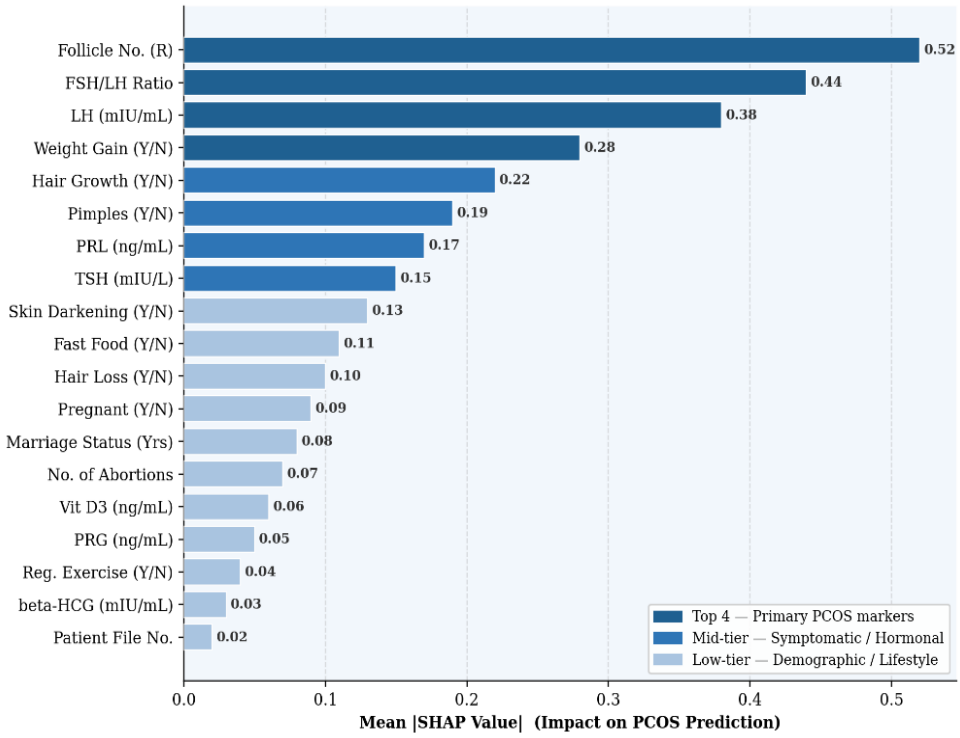

Fig. 8 SHAP Bar Importance Graph — mean absolute SHAP value per feature, grouped into primary PCOS markers, symptomatic/hormonal predictors, and demographic/lifestyle features.

5. CONCLUSION

This study proposed a novel, reproducible machine-learning pipeline for PCOS detection that departs from existing survey-based efforts through three methods uncommon in the literature: iterative Variance Inflation Factor (VIF) removal to eliminate multicollinearity, SMOTE-based class balancing, and SHAP-based explainability applied to the most effective model. Using a merged multicentric dataset of 541 patient records collected across ten Kerala hospitals, the proposed pipeline selected 19 multicollinearity-free predictors (all VIF< 5) out of 41 raw hormonal and clinical features through ten rounds of successive elimination, ensuring that the remaining features — such as LH (VIF 1.03), the FSH/LH ratio (VIF 1.03), and Follicle No. (R) (VIF 3.58) — contribute only genuinely independent predictive variance to the model.

Class balancing, applied only to the training set, corrected the 2:1 class imbalance (252 PCOS-negative vs. 125 PCOS-positive) by synthetically increasing the minority class to 252 samples. Comparison with the scale_pos_weight alternative showed that SMOTE reduced false positives from eleven to eight while maintaining identical false negatives (eight), raising overall accuracy from 88.27% to 90.12%. The final tuned XGBoost model (n_estimators=700, learning_rate=0.02, gamma=0.1, min_child_weight=4) trained on the SMOTE-balanced, VIF-reduced feature set achieved the best overall performance: 90.12% accuracy, 84% PCOS-positive precision and recall, and an F1-score of 0.84 on the 162-record held-out test set, correctly identifying 43 of 51 true PCOS-positive cases with only eight false negatives.

SHAP analysis of the XGBoost model on the test data identified Follicle No. (R), the FSH/LH ratio, LH, weight gain, and androgenic symptoms as the top predictive variables, mapping directly onto the Rotterdam diagnostic criteria for PCOS. The negative SHAP contributions of TSH and prolactin indicate that the model learned to differentiate PCOS from thyroid dysfunction and hyperprolactinaemia — a clinically significant differential-diagnosis capability that emerged naturally from the training data. Taken together, these findings provide a validated, multicollinearity-adjusted and interpretable baseline for PCOS screening research, offering a methodological foundation that future studies could extend toward prospective clinical validation, multi-site generalisation, and integration with real-time electronic health record systems.

References

1. Kottarathil P, Antony Belcy M, Kovoor P, Dev S. PCOS dataset [Data set]. Kaggle. 2020. Available from: kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos
2. Kyrou I, Randeva HS, Tsigos C, Kaltsas G, Weickert MO. Clinical problems caused by obesity. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000. PMID: 25905207.
3. Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Human Reproduction. 2016;31(12):2841–2855.
4. Teede HJ, Misso ML, Costello MF, Dokras A, Laven J, Moran L, et al. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Human Reproduction. 2018;33(9):1602–1618.
5. Barrera FJ, Sheng B, Mehta A, Bhatt DL, Koh JM. Machine learning in the diagnosis and management of polycystic ovary syndrome: a systematic review. Frontiers in Endocrinology. 2023;14:1106728.
6. Uddin MF, Hossain MA, Hasan MM, Reza T, Ahmed S. Early detection of polycystic ovary syndrome: a comparative analysis of traditional and ensemble machine learning algorithms with explainability. Engineering Reports. 2025;7(3):e12942.
7. Rahman MA, Hossen A, Karim MR, Mridha MF. PCOS detection using machine learning classifiers with clinical hormonal markers. Informatics in Medicine Unlocked. 2024;46:101469.
8. Tong C, Wu J, Lin Q, Zheng Y, Zhang Y. A diagnostic model for polycystic ovary syndrome based on machine learning. Scientific Reports. 2025;15:9821.
9. Elmannai H, El-Rashidy N, Mashal I, Alohali MA, Farag MM, El-Sappagh S, et al. Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. Diagnostics. 2023;13(7):1304.
10. Panjwani M, Ahmad I, Anwar MS, Imran M, Alruwaili M. Optimized machine learning for enhanced PCOS risk prediction using clinical data. Sensors. 2025;25(3):711.
11. Zhang Y, Chen L, Wang F, Liu Y, Zhao M. Explainable machine learning and nomogram model for early detection of polycystic ovary syndrome: a multicenter study. Frontiers in Endocrinology. 2025;16:1719631.
12. Emara HM, Shouman M, El-Fishawy N, Khalifa NEM. A stacked learning framework with BORUTA feature selection and SMOTE for imbalanced PCOS classification. Frontiers in Endocrinology. 2025;16:1578342.
13. Gomes MO, Barra NG, Ferreira C, Gomes CE, Fontaine V, Nadeau AJ, et al. Anti-Müllerian hormone as a diagnostic marker of polycystic ovary syndrome: a systematic review and meta-analysis. American Journal of Obstetrics and Gynecology. 2025;232(5):506–523.
14. Yue CY, Lu LC, Li M, Zhang QM, Ying CM. Threshold value of anti-Müllerian hormone for the diagnosis of polycystic ovary syndrome in Chinese women. Scientific Reports. 2018;8(1):9389.
15. Zad AS, Mirheidari N, Soleymani F, Tabrizi M, Farokhimanesh S. Machine learning-based diagnosis of polycystic ovary syndrome using large-scale electronic health records. Frontiers in Endocrinology. 2024;14:1289549.
16. Karimi-Haghighi S, Mousavinia SEH, Rezaei MH, Sharafkhah A. Non-invasive machine learning approach for PCOS diagnosis using ultrasound and clinical features with SHAP explainability. Scientific Reports. 2025;15:12847.
17. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R. 2nd ed. New York: Springer; 2021.
18. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321–357.
19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA. New York: ACM; 2016. p. 785–794.
20. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proc. 32nd Conf. Neural Information Processing Systems (NeurIPS); 2018 Dec 3–8; Montréal, Canada. Curran Associates; 2018. p. 6638–6648.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proc. 31st Conf. Neural Information Processing Systems (NeurIPS); 2017 Dec 4–9; Long Beach, CA. Curran Associates; 2017. p. 4768–4777.
22. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertility and Sterility. 2004;81(1):19–25.
23. La Radiologia Medica. Artificial intelligence in polycystic ovary syndrome management: past, present, and future — a comprehensive review. La Radiologia Medica. 2025;130(2):214–231.
24. Bisht N, Tripathi MC, Tiwari VK. Comparative analytic study of machine learning algorithms for the diagnosis of polycystic ovary syndrome. In: Proc. 5th Int. Conf. Intelligent Computing and Research Advancement (ICIRCA); 2023 Jun 1–3; Coimbatore, India. IEEE; 2023.
25. Saveetha Engineering College. PCOS detection using machine learning and medical app development. In: Proc. Int. Conf. Sustainable Technologies for Energy and Environmental Development (SUSTAINED); 2024. IEEE; 2024.
26. Nimmala B, Indrakanti NR, Katepalli R, Reddy VK. PCOS detection and real-time monitoring using machine learning-integrated mobile health application. In: Proc. Int. Conf. Intelligent Processing and Communication Networks (ICIPCN); 2024; Kathmandu, Nepal. IEEE; 2024.
27. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13(1):21–27.
28. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.

29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.