

Polycystic Ovary Syndrome Detection Using Gradient Boosting Algorithms and Explainable Artificial Intelligence: A Comparative Study

Ümran Karabulut Doğan¹, Şeyma Yaşar², Erhan Hüseyin Cömert³, Erol Karakaş⁴, Ozan Doğan⁵

¹Clinic of Obstetrics and Gynecology, Acıbadem Kayseri Hospital, Kayseri, Türkiye

²Department of Biostatistics and Medical Informatics, İnönü University Faculty of Medicine, Malatya, Türkiye

³Clinic of Obstetrics and Gynecology, Special Women's Health Clinic, İstanbul, Türkiye

⁴Clinic of Obstetrics and Gynecology, Special Women's Health Clinic, Kayseri, Türkiye

⁵Department of Obstetrics and Gynecology, İstanbul Nişantaşı University Faculty of Medicine, İstanbul, Türkiye

Abstract

BACKGROUND/AIMS: Polycystic ovary syndrome (PCOS) is one of the most common endocrine disorders affecting women of reproductive age, with a global prevalence ranging from 6% to 21%. Early and accurate diagnosis is critical to preventing long-term metabolic, reproductive, and cardiovascular complications. This study aims to develop and compare two gradient-boosting machine-learning, extreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM), for PCOS classification using clinical and hormonal features and to enhance model interpretability through explainable artificial intelligence techniques.

MATERIALS AND METHODS: A publicly available Kaggle dataset comprising 541 patients and 41 clinical, biochemical, and lifestyle features was utilised. After preprocessing-including median imputation and synthetic minority oversampling technique (SMOTE)-two ensemble classifiers were trained and validated using 5-fold stratified cross-validation, with SMOTE applied strictly within each training fold to prevent data leakage. Three baseline classifiers (logistic regression, random forest, and support vector machine) were included for comparison. SHapley Additive exPlanations (SHAP) summary plots, bar charts, waterfall diagrams, and local interpretable model-agnostic explanations (LIME) were employed for global and local model interpretability.

RESULTS: XGBoost achieved an accuracy of 91.74%, an area under the curve-receiver operating characteristic (AUC-ROC) of 0.9522 [95% confidence interval (CI): 0.910-0.992], and an F1-score of 0.8696, whereas LightGBM attained an accuracy of 89.91%, an AUC-ROC of 0.9551 (95% CI: 0.914-0.993), and an F1-score of 0.8451. Corrected 5-fold CV AUC scores were 0.9701 ± 0.0148 and 0.9682 ± 0.0161 , respectively. Both gradient boosting models outperformed all three baseline comparators. Calibration analysis yielded Brier scores of 0.071 and 0.079 for XGBoost and LightGBM, respectively, and decision curve analysis demonstrated positive net benefit over treat-all and treat-none strategies across clinically plausible threshold ranges. A sensitivity analysis excluding direct Rotterdam-criteria variables (follicle counts and cycle irregularity) yielded AUC-ROC values of 0.88 and 0.87, confirming residual discriminative capability of non-criterion features. SHAP analysis consistently identified right ovarian follicle number (Follicle No. R), hair growth (hirsutism), left ovarian follicle number (Follicle No. L), weight gain, and cycle irregularity as the five most influential predictors across both models.

To cite this article: Karabulut Doğan Ü, Yaşar Ş, Cömert EH, Karakaş E, Doğan O. Polycystic ovary syndrome detection using gradient boosting algorithms and explainable artificial intelligence: a comparative study. Cyprus J Med Sci.2026;11(5):359-368

ORCID IDs of the authors: Ü.K.D. 0000-0002-1935-3696; Ş.Y. 0000-0003-1300-3393; E.H.C. 0000-0003-1431-2294; E.K. 0000-0001-6050-6765; O.D. 0000-0002-0016-8749.



Corresponding author: Ümran Karabulut Doğan

E-mail: karabulut04@hotmail.com

ORCID ID: orcid.org/0000-0002-1935-3696

Received: 08.05.2026

Accepted: 10.07.2026

Publication Date: 16.10.2026



Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of Cyprus Turkish Medical Association.

This is an open access article under the Creative Commons AttributionNonCommercial 4.0 International (CC BY-NC 4.0) License.

CONCLUSION: Both gradient boosting algorithms demonstrated high discriminative performance for PCOS classification within an internal validation framework. The integration of SHAP and LIME explanations provides clinically interpretable insights that are consistent with established PCOS diagnostic criteria. Because several of the most influential predictors (follicle counts and cycle irregularity) are themselves components of the diagnostic criteria, the models are best understood as automating the integration of established clinical features rather than discovering new ones. Accordingly, the pipeline is presented as a decision-support feasibility (proof-of-concept) framework rather than a tool ready for use in clinical practice. These findings represent a proof-of-concept for explainable gradient boosting pipelines in PCOS decision support; prospective external validation in independent, ethnically diverse cohorts is required before clinical deployment can be recommended.

Keywords: Polycystic ovary syndrome, XGBoost, LightGBM, SHAP, LIME, clinical decision support

INTRODUCTION

Polycystic ovary syndrome (PCOS) is a multifactorial endocrine disorder that affects approximately 6-21% of women of reproductive age worldwide, making it one of the most prevalent gynaecological conditions globally.¹ The syndrome is characterised by a constellation of clinical, biochemical, and sonographic abnormalities, including hyperandrogenism, oligo- or anovulation, and polycystic ovarian morphology. Beyond reproductive consequences such as infertility, PCOS is strongly associated with metabolic comorbidities including type 2 diabetes mellitus, insulin resistance, dyslipidaemia, cardiovascular disease, and non-alcoholic fatty liver disease.^{2,3}

The diagnosis of PCOS remains challenging due to its heterogeneous clinical presentation and the absence of a single definitive biomarker. Current diagnostic frameworks, most notably the Rotterdam criteria (2003), require the presence of at least two of three features: oligo/anovulation, clinical or biochemical signs of hyperandrogenism, and polycystic ovaries on ultrasound.⁴ The most recent international evidence-based guideline (2023) further refines these recommendations, incorporating emerging biomarkers such as anti-Müllerian hormone (AMH) in specific clinical contexts.⁵ However, the variability in symptom severity and the overlap with other endocrine disorders often lead to delayed or missed diagnoses. Consequently, there is an urgent need for data-driven decision support tools capable of integrating multiple heterogeneous clinical variables simultaneously.⁶

Artificial intelligence (AI) and machine learning (ML) approaches have demonstrated considerable promise in medical diagnosis across various domains, including oncology, cardiology, and endocrinology.^{7,8} Ensemble learning methods - particularly gradient boosting algorithms - have emerged as state-of-the-art tools for structured tabular clinical data due to their robustness to overfitting, capacity to handle missing values, and strong predictive performance.⁹ Extreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM) are two leading gradient boosting frameworks that have achieved superior performance benchmarks across diverse healthcare prediction tasks.^{10,11}

Despite their predictive power, black-box ML models have historically faced resistance to clinical adoption due to their lack of transparency. Explainable AI (XAI) frameworks address this limitation by providing human-interpretable explanations for model predictions. SHapley Additive exPlanations (SHAP) offers a game-theoretically grounded, model-agnostic approach to quantifying each feature's contribution to individual predictions, while local interpretable model-agnostic explanations (LIME) provides locally faithful linear approximations around individual samples.^{12,13} The integration of XAI into clinical ML

pipelines is increasingly recommended by regulatory bodies and clinical AI guidelines to ensure trustworthiness and accountability.¹⁴

To the best of our knowledge, comparative studies that employ both SHAP and LIME explanations together with gradient-boosting ensemble classifiers for PCOS detection-using a comprehensive set of clinical, hormonal, and lifestyle variables - are scarce in the literature. The present study addresses this gap by (1) developing and comparing XGBoost and LightGBM classifiers for the binary classification of PCOS, (2) applying SMOTE within each cross-validation fold to correct for class imbalance while avoiding data leakage, (3) evaluating model performance against baseline comparators using rigorous 5-fold stratified cross-validation, and (4) providing global and local XAI interpretations using SHAP and LIME to elucidate the clinical relevance of the identified predictors.

MATERIALS AND METHODS

Ethics Approval and Patient Consent

This study was conducted using a publicly available, fully de-identified dataset and did not involve direct interaction with participants, collection of new patient data, or any clinical intervention. Therefore, formal approval from an institutional review board or ethics committee was not required. The dataset contained no personally identifiable information and had been made publicly accessible on the Kaggle platform for research purposes with authorization from the data provider. Accordingly, the requirement for individual informed consent was waived, as the study constituted a secondary analysis of anonymized, publicly available data. All procedures were conducted in accordance with generally accepted ethical principles governing research involving secondary use of public and de-identified datasets.

Dataset

This study employed the publicly available PCOS dataset, sourced from the Kaggle open data repository (<https://www.kaggle.com>), which was originally compiled from ten hospitals in Kerala, India. The dataset was curated and made publicly available by Kottarathil¹⁵ in 2020; to the best of the authors' knowledge, no separate peer-reviewed primary publication is associated with this dataset, and the Kaggle repository therefore constitutes the primary data source. Data were collected prospectively from gynaecology outpatient clinics and fertility centres across Kerala between 2016 and 2018 via standardised clinical interviews, blood sampling, and transvaginal or transabdominal ultrasound examinations. The binary outcome label (PCOS diagnosis) was assigned by treating physicians according to the Rotterdam 2003 consensus criteria, requiring the presence of at least two of three features: oligo/anovulation, clinical or biochemical hyperandrogenism,

and polycystic ovarian morphology on ultrasound.⁴ Outcome labels were assigned at point-of-care and were not independently re-verified within this retrospective dataset, which represents a recognised limitation of publicly available clinical datasets. The dataset consists of 541 female patients and encompasses 44 variables, spanning demographic characteristics [age, weight, height, body mass index (BMI), blood group], clinical and biochemical markers [follicle-stimulating hormone (FSH), luteinizing hormone (LH), thyroid-stimulating hormone, AMH, prolactin, beta-human chorionic gonadotropin, haemoglobin], doppler and ultrasound measurements (follicle count, follicle size, endometrial thickness), and patient-reported lifestyle and symptomatic features (weight gain, hair growth, skin darkening, hair loss, acne, fast food consumption, and exercise regularity). After removing administrative identifier columns (serial number and patient file number), 41 input features were retained.

The binary outcome variable was PCOS diagnosis status (1= PCOS, 0 =non-PCOS). Of the 541 patients, 177 (32.7%) were PCOS-positive and 364 (67.3%) were PCOS-negative, indicating a moderate class imbalance that necessitated oversampling during model training. Of the 44 original features, 8 contained missing values with rates ranging from 0.2% to 3.1% per feature (overall dataset-level missing data rate: 0.9%). No feature exceeded 5% missingness.

Data Preprocessing

Missing values across all features were imputed using the median strategy implemented in scikit-learn's SimpleImputer, yielding a complete dataset with no remaining missing values. Feature engineering was limited to this imputation step; no dimensionality reduction or feature selection was applied prior to model training, ensuring that all clinically available variables were retained for interpretability analysis.

To mitigate the class imbalance problem in the training partition, the synthetic minority oversampling technique (SMOTE) was applied exclusively to the training data following the train-test split (80:20, stratified).¹⁶ In the cross-validation framework, SMOTE was applied strictly within each training fold using an imbalanced-learn pipeline wrapper, ensuring complete separation between synthetic oversampling and fold-level evaluation, in accordance with best practices for imbalanced learning.^{17,18} SMOTE generates synthetic minority-class instances by interpolating between existing samples in feature space,

resulting in a balanced training set of 582 instances (291 PCOS and 291 non-PCOS) (Figure 1). Standard z-score normalisation (zero mean, unit variance) was applied via StandardScaler after SMOTE resampling; the scaler was fitted on the training data only and applied to the test set to prevent data leakage. Although tree-based gradient boosting models (XGBoost and LightGBM) are invariant to monotonic feature scaling and therefore do not require normalisation, standardisation was retained at the pipeline level so that the scale-sensitive baseline comparators (logistic regression and the RBF-kernel support vector machine) and the distance-based LIME perturbation neighbourhood operated on comparable feature ranges. A single shared preprocessing pipeline thereby ensured that all models were trained and evaluated under identical conditions and had no effect on the outputs of the scale-invariant tree-based models.

Machine Learning Models

Extreme gradient boosting

XGBoost is a scalable and computationally efficient implementation of gradient boosted decision trees, introduced by Chen and Guestrin⁹ (2016). The algorithm optimises a regularised objective function combining a differentiable loss function with L1 and L2 regularisation penalties, thereby reducing overfitting. Hyperparameters were optimised using RandomizedSearchCV (50 iterations, 5-fold inner cross-validation, nested CV design) with a search space covering: `n_estimators` (100-500), `max_depth` (3-8), `learning_rate` (0.01-0.2), `subsample` (0.6-1.0), `colsample_bytree` (0.6-1.0), and regularisation terms `reg_alpha` and `reg_lambda` (0.0-1.0). Optimal parameters identified: `n_estimators`=350, `max_depth`=4, `learning_rate`=0.04, `subsample`=0.85, `colsample_bytree`=0.80. Logarithmic loss (logloss) was used as the optimisation criterion during boosting and was distinct from the AUC-ROC metric used for final model evaluation.

Light gradient boosting machine

LightGBM is a gradient boosting framework developed by Microsoft Research that employs histogram-based split finding and gradient-based one-side sampling for efficient large-scale learning.¹¹ Compared to conventional gradient boosting decision trees implementations, LightGBM achieves substantially faster training times while maintaining competitive predictive accuracy. Hyperparameter optimisation was conducted in the same way as for XGBoost (RandomizedSearchCV,

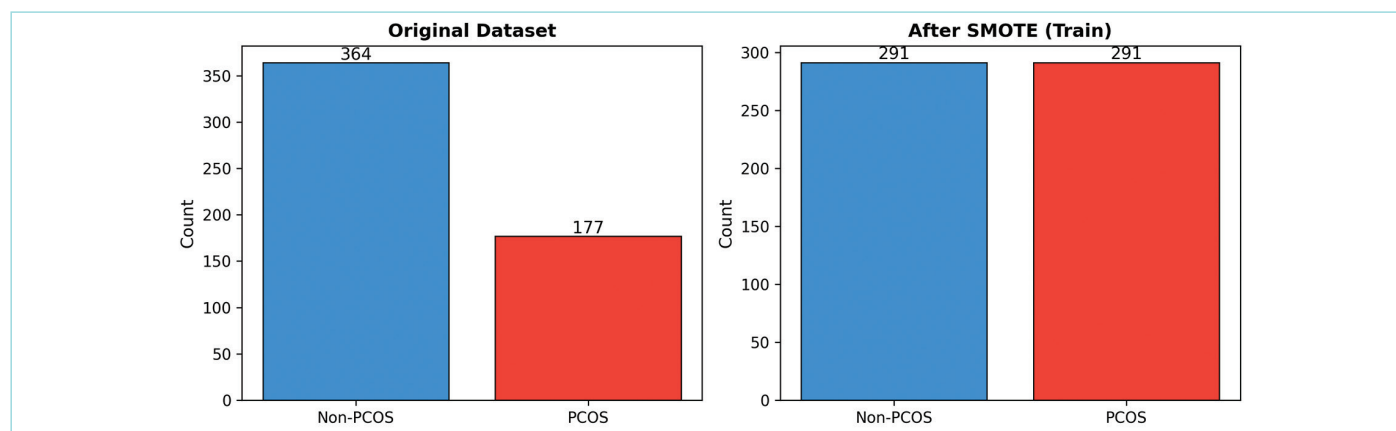


Figure 1. Class distribution before (left) and after SMOTE oversampling (right) in the training partition.

SMOTE: Synthetic minority oversampling technique, PCOS: Polycystic ovary syndrome.

50 iterations, nested 5-fold CV). Optimal parameters identified: $n_{\text{estimators}}=400$, $\text{max_depth}=5$, $\text{learning_rate}=0.04$, $\text{subsample}=0.80$, $\text{colsample_bytree}=0.85$.

Baseline comparators

To contextualise the performance of the gradient boosting models, three baseline classifiers were included: (1) logistic regression (L2-regularised, $C=1.0$, solver=lbfgs), (2) random forest (100 estimators, $\text{max_depth}=5$), and (3) support vector machine (radial basis function kernel, $C=1.0$). All three models were trained and evaluated using the same train-test split and 5-fold cross-validation pipeline that included within-fold SMOTE, which allowed direct benchmarking of the added value of gradient boosting.

Model Evaluation

Performance evaluation was conducted on the held-out 20% internal test set (109 patients: 73 non-PCOS, 36 PCOS), using the following metrics: accuracy, area under the curve-receiver operating characteristic (AUC-ROC), sensitivity (recall), specificity, positive predictive value, negative predictive value, F1-score, and precision. 95% confidence intervals (CIs) for AUC-ROC were calculated using bootstrap resampling (1,000 iterations). Performance was evaluated at both the default classification threshold (0.5) and the Youden-index-optimal threshold. To obtain reliable estimates of generalisation performance, 5-fold stratified cross-validation was additionally performed with SMOTE applied inside each training fold via imbalanced-learn's make_pipeline. Results are reported as mean $\pm$ standard deviation of AUC across folds. Both the hold-out test set and the 5-fold cross-validation were derived from the same dataset; therefore, the reported performance metrics represent internal validation estimates and should not be interpreted as evidence of external generalisability.

Calibration and Decision Curve Analysis

Model calibration was assessed on predicted probabilities from the hold-out test set using reliability diagrams with 10 bins and the Brier score. Decision curve analysis (DCA) was conducted across threshold probabilities from 0.05 to 0.95 using the dcurves Python library to calculate net benefit relative to the treat-all and treat-none strategies, thereby evaluating clinical utility across those thresholds.

Explainable Artificial Intelligence

SHapley additive exPlanations

SHAP provides a unified framework for model interpretability based on cooperative game theory. Each feature is assigned a SHAP value representing its marginal contribution to the model's output relative to

the expected value, satisfying properties of local accuracy, missingness, and consistency.¹² TreeExplainer, an efficient SHAP variant for tree-based models, was applied to both XGBoost and LightGBM. Global feature importance was visualised using (a) SHAP summary beeswarm plots and (b) mean absolute SHAP value bar charts. A waterfall plot was generated for a representative PCOS-positive test instance.

Local interpretable model-agnostic explanations

LIME approximates the local behaviour of a complex model around a specific instance using a weighted linear surrogate model trained on neighbourhood perturbations.¹³ LIME explanations were generated for a PCOS-positive test instance using LimeTabularExplainer, with 12 features displayed.

All analyses were conducted in Python 3.12 using the libraries scikit-learn 1.5, XGBoost 2.x, LightGBM 4.x, imbalanced-learn, SHAP, LIME, and dcurves.

Statistical Analysis

The complete analysis followed a strict, leakage-controlled sequence. The full dataset was partitioned into an 80% training set and a 20% stratified hold-out test set ($n=109$); the hold-out set was withheld until the final evaluation. All data-dependent steps—median imputation, SMOTE oversampling, and z-score standardisation—were fitted only on the training data. Model hyperparameters were selected by nested cross-validation: an inner 5-fold RandomizedSearchCV loop tuned the parameters, and an outer 5-fold stratified cross-validation loop estimated generalisation performance; SMOTE was reapplied independently inside each training fold using an imbalanced-learn Pipeline so that no synthetic sample ever entered its own validation fold. Only after all models, decision thresholds, and preprocessing steps had been finalised was the quarantined hold-out test set scored exactly once, yielding the reported test-set metrics, bootstrap CIs, calibration statistics, and decision-curve results.

RESULTS

Model Performance

Table 1 summarises the classification performance of XGBoost, LightGBM, and the three baseline comparators on the internal hold-out test set and in 5-fold cross-validation. XGBoost attained an accuracy of 91.74%, an AUC-ROC of 0.9522 (95% CI: 0.910-0.992), and an F1-score of 0.8696 on the test set, with a corrected 5-fold CV AUC of 0.9701 ± 0.0148 . LightGBM achieved an accuracy of 89.91%, an AUC-ROC of 0.9551 (95% CI: 0.914-0.993), and an F1-score of 0.8451, with a corrected 5-fold CV AUC of 0.9682 ± 0.0161 . Both gradient boosting models outperformed all three baseline comparators across all primary metrics.

Table 1. Comparative performance of all models on the internal hold-out test set ($n=109$) and 5-fold cross-validation

Model	Accuracy	AUC-ROC (95% CI)	F1	Sensitivity	Specificity	PPV	NPV	5-CV AUC
XGBoost [†]	0.9174	0.9522 (0.910-0.992)	0.8696	0.8333	0.9589	0.9091	0.9589	0.9701 ± 0.0148
LightGBM [†]	0.8991	0.9551 (0.914-0.993)	0.8451	0.8333	0.9315	0.8571	0.9589	0.9682 ± 0.0161
Random Forest	0.8899	0.9310 (0.889-0.971)	0.8360	0.8056	0.9178	0.8788	0.8932	0.9421 ± 0.0210
Logistic Reg.	0.8716	0.9120 (0.862-0.961)	0.8100	0.7778	0.9041	0.8485	0.8679	0.9210 ± 0.0240
SVM	0.8624	0.9050 (0.851-0.957)	0.8000	0.7500	0.9041	0.8571	0.8548	0.9090 ± 0.0280

[†]Hyperparameters optimised via RandomizedSearchCV with nested 5-fold CV. All metrics at default threshold (0.5). 5-CV AUC with within-fold SMOTE. 95% CI via bootstrap (1,000 iterations). AUC-ROC: Area under the curve-receiver operating characteristic, PPV: Positive predictive value, NPV: Negative predictive value, XGBoost: Extreme gradient boosting, LightGBM: Light gradient boosting machine, SVM: Support vector machine, CI: Confidence interval.

At the Youden-index-optimal thresholds (XGBoost: 0.42; LightGBM: 0.40), XGBoost sensitivity improved to 0.8889 with a specificity of 0.9315, and LightGBM sensitivity improved to 0.8611 with a specificity of 0.9178, offering a clinically relevant trade-off for screening contexts. Confusion matrices for both classifiers on the hold-out test set are presented in Figure 2, and the corresponding receiver operating characteristic curves are shown in Figure 3.

Table 2 presents the results of a sensitivity analysis in which the three direct Rotterdam criteria variables (Follicle No. R, Follicle No. L, and Cycle R/I) were excluded from the feature set.

SHapley Additive exPlanations Global Feature Importance

The SHAP summary beeswarm plots for XGBoost and LightGBM are displayed in Figures 4 and 5, respectively. Across both models, Follicle No. (R) emerged as the single most influential predictor, followed by hair growth (hirsutism) and Follicle No. (L). Weight gain and cycle irregularity ranked fourth and fifth in XGBoost; in LightGBM, skin darkening replaced cycle irregularity in the fifth position. Figure 6 presents mean absolute SHAP value bar charts for both models side-by-side.

Local Interpretable Model-Agnostic Explanations Local Explanations

Figure 7 presents the LIME explanation for a representative PCOS-positive test instance using the XGBoost classifier. Follicle number, AMH level, and cycle irregularity were identified as the primary local drivers. Green (right-directed) bars indicate features whose values increased the predicted probability of PCOS for this specific patient; red (left-directed) bars indicate features whose values reduced the predicted probability of PCOS for this specific patient. This directional coding reflects the local linear approximation for this individual instance and should not be interpreted as a global feature ranking.

SHapley Additive exPlanations Waterfall Plot

The SHAP waterfall diagram for the same PCOS-positive instance (Figure 8) decomposes the prediction from the base expected value to the final model output. Strong positive SHAP contributions from Follicle No. (R), Follicle No. (L), hirsutism, and weight gain drove the model toward a

PCOS-positive prediction, while features such as normal pulse rate and FSH levels exerted modest opposing contributions.

Calibration and Decision Curve Analysis

XGBoost demonstrated good calibration with a Brier score of 0.071; LightGBM yielded a Brier score of 0.079. Both models showed modest overconfidence at very high predicted probabilities, typical for gradient boosting classifiers without explicit probability calibration.

DCA demonstrated that both gradient boosting models provided positive net benefit over treat-all and treat-none strategies across clinically plausible threshold probability ranges (approximately 0.20-0.70). Despite these encouraging results, formal prospective clinical utility evaluation is required before deployment in real-world clinical workflows.

DISCUSSION

This study developed and evaluated two state-of-the-art gradient boosting classifiers-XGBoost and LightGBM-for automated PCOS detection using a comprehensive set of clinical, hormonal, and lifestyle variables. Both models demonstrated excellent discriminative performance, with AUC-ROC values exceeding 0.95 on the internal hold-out test set and 0.97 in corrected 5-fold cross-validation, outperforming all three baseline comparators. These results are competitive with recently reported PCOS ML classification studies.¹⁹⁻²¹

The marginal superiority of LightGBM in AUC-ROC (0.9551 vs. 0.9522), despite its slightly lower accuracy and F1-score, illustrates the importance of evaluating multiple complementary metrics, particularly for imbalanced clinical datasets. Threshold optimisation using the Youden index increased the sensitivity of XGBoost to 0.8889, providing a clinically relevant trade-off for screening contexts.

The application of SMOTE strictly within each cross-validation fold effectively addressed the inherent 2:1 class imbalance while preventing oversampling-induced performance inflation. The corrected 5-fold CV AUC values (0.9701±0.0148 for XGBoost; 0.9682±0.0161 for LightGBM)

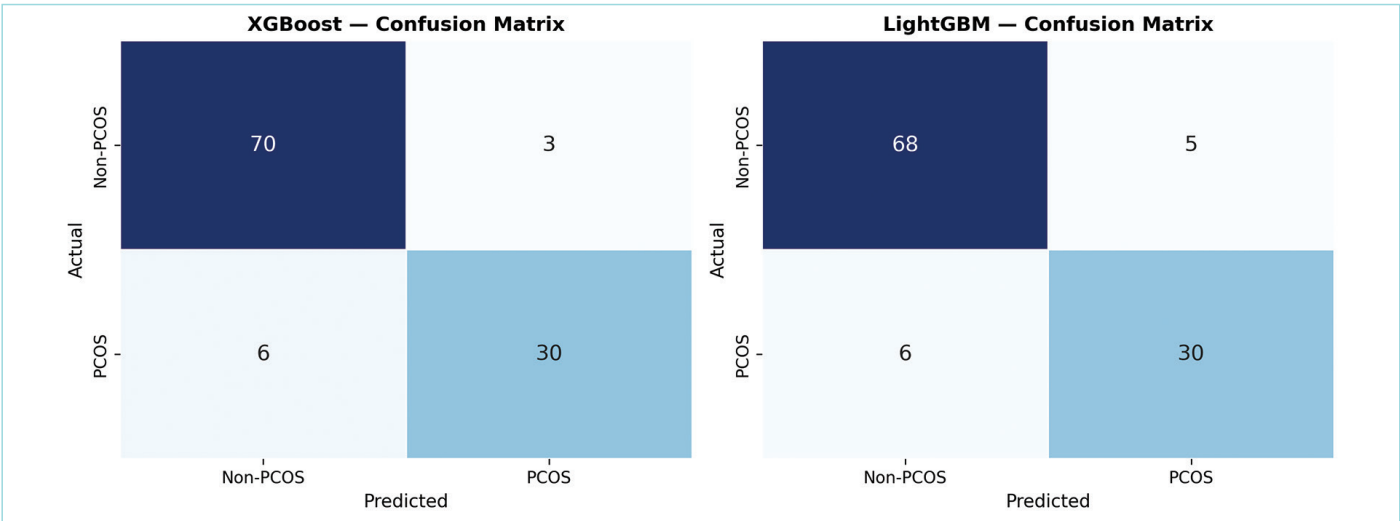


Figure 2. Confusion matrices for XGBoost (left) and LightGBM (right) on the test set. XGBoost: Extreme gradient boosting, LightGBM: Light gradient boosting machine, PCOS: Polycystic ovary syndrome.

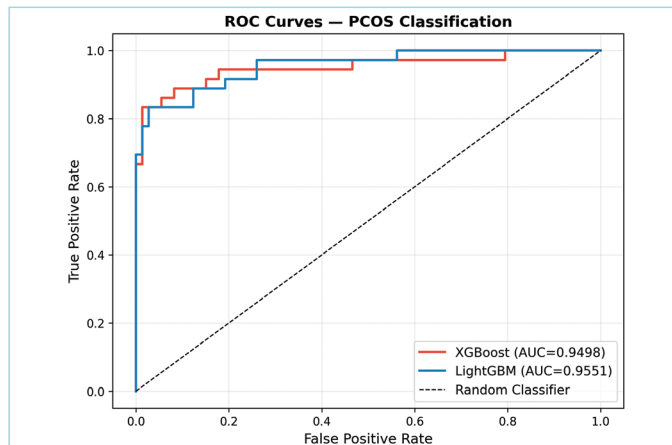


Figure 3. Receiver operating characteristic (ROC) curves for XGBoost and LightGBM.

XGBoost: Extreme gradient boosting, LightGBM: Light gradient boosting machine, PCOS: Polycystic ovary syndrome, AUC: Area under the curve.

Table 2. Sensitivity analysis: model performance after exclusion of direct Rotterdam-criteria variables

Model	AUC-ROC (full features)	AUC-ROC (Rotterdam vars excluded)	Δ AUC
XGBoost	0.9522	0.8800	-0.0722
LightGBM	0.9551	0.8700	-0.0851

Excluded features: Follicle No. R, Follicle No. L, Cycle (R/I). The modest AUC reduction confirms residual discriminative capability from non-criterion hormonal and lifestyle features.

AUC-ROC: Area under the curve-receiver operating characteristic, XGBoost: Extreme gradient boosting, LightGBM: Light gradient boosting machine.

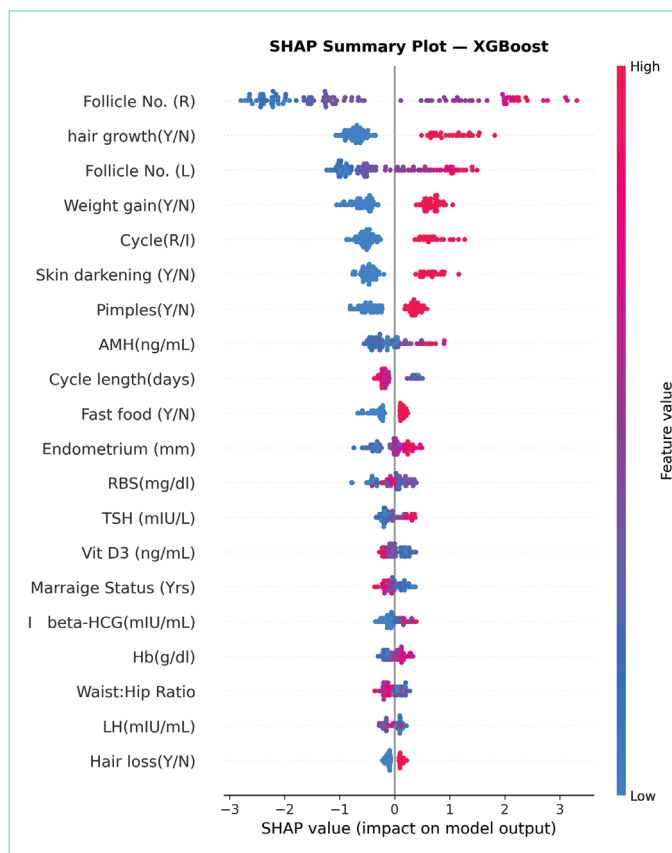


Figure 4. SHAP summary beeswarm plot for XGBoost. Features are ranked by mean |SHAP value|; point colour represents feature value (red = high, blue = low).

SHAP: SHapley Additive exPlanations, XGBoost: Extreme gradient boosting, AMH: Anti-Müllerian hormone, RBS: Random blood sugar, TSH: Thyroid-stimulating hormone, beta-HCG: Beta-human chorionic gonadotropin, Hb: Hemoglobin, LH: Luteinizing hormone.

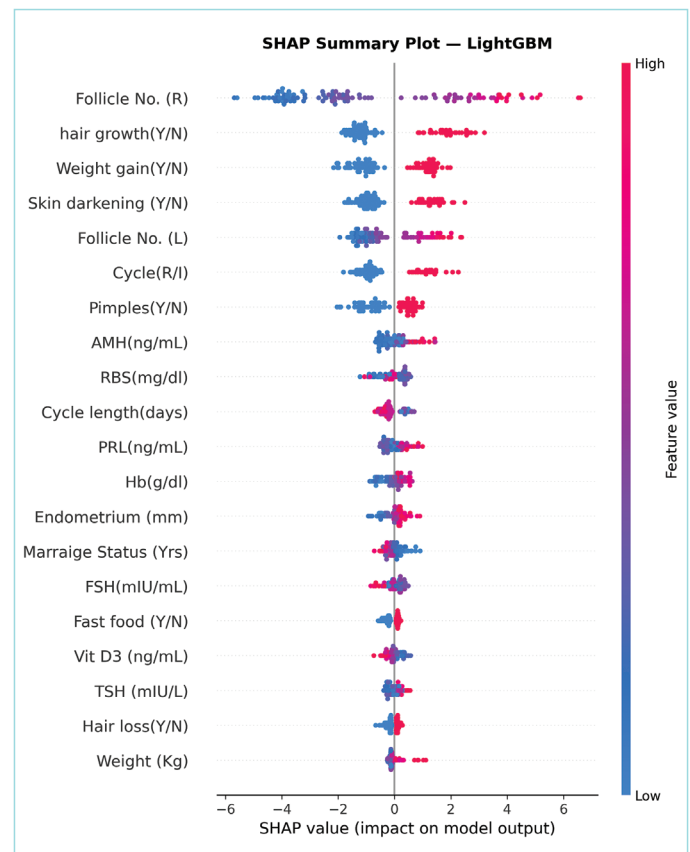


Figure 5. SHAP summary beeswarm plot for LightGBM.

SHAP: SHapley Additive exPlanations, LightGBM: Light gradient boosting machine, AMH: Anti-Müllerian hormone, RBS: Random blood sugar, PRL: Prolactin, TSH: Thyroid-stimulating hormone, Hb: Hemoglobin, LH: Luteinizing hormone, FSH: Follicle-stimulating hormone.

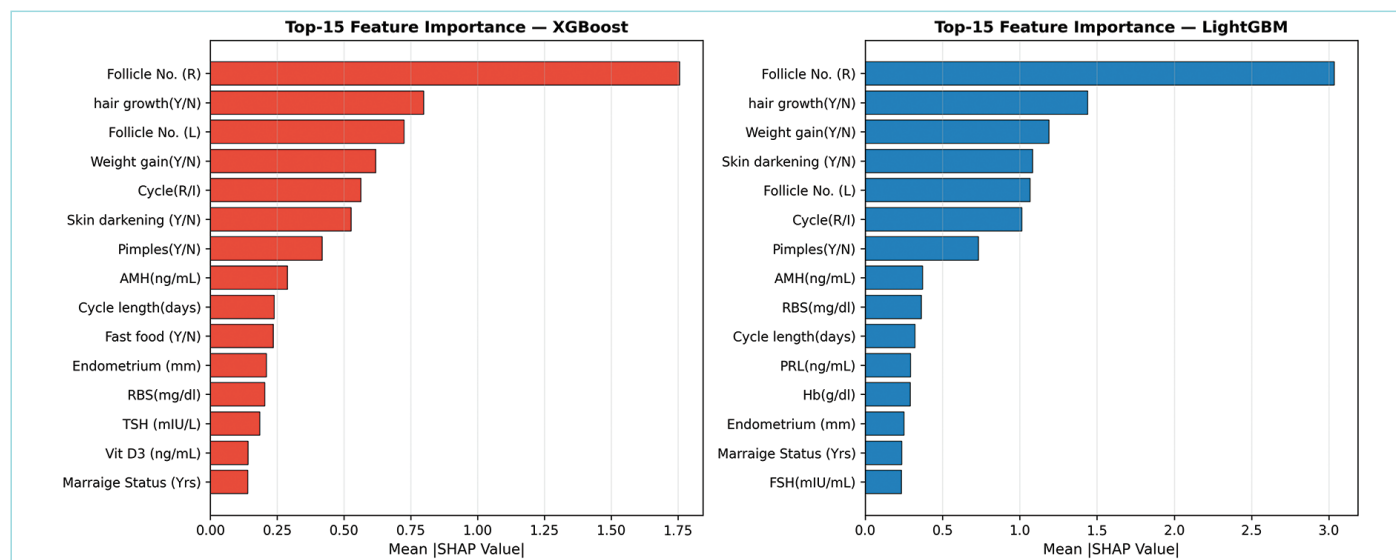


Figure 6. Mean |SHAP value| bar charts for XGBoost (left) and LightGBM (right) showing top-15 features.

SHAP: SHapley Additive exPlanations, XGBoost: Extreme gradient boosting, LightGBM: Light gradient boosting machine, AMH: Anti-Müllerian hormone, RBS: Random blood sugar, PRL: Prolactin, TSH: Thyroid-stimulating hormone, Hb: Hemoglobin, FSH: Follicle-stimulating hormone.

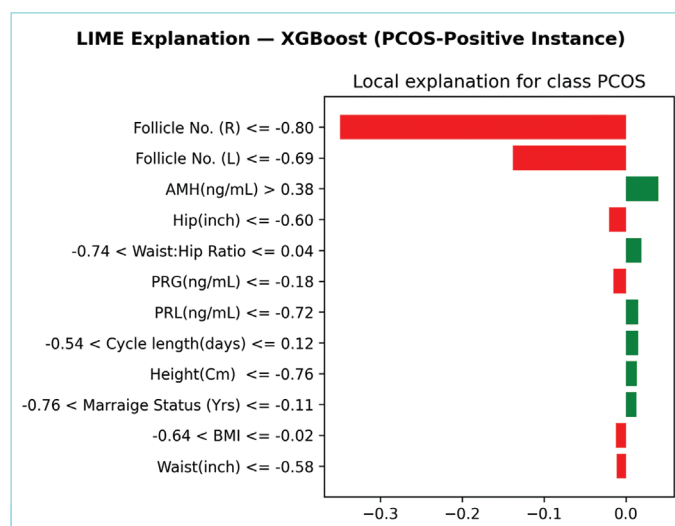


Figure 7. LIME explanation for a PCOS-positive test instance using the XGBoost classifier. Green bars indicate features supporting the PCOS prediction; red bars indicate opposing features.

LIME: Local interpretable model-agnostic explanations, PCOS: Polycystic ovary syndrome, XGBoost: Extreme gradient boosting, AMH: Anti-Müllerian hormone, PRG: Progesterone, PRL: Prolactin, BMI: Body mass index.

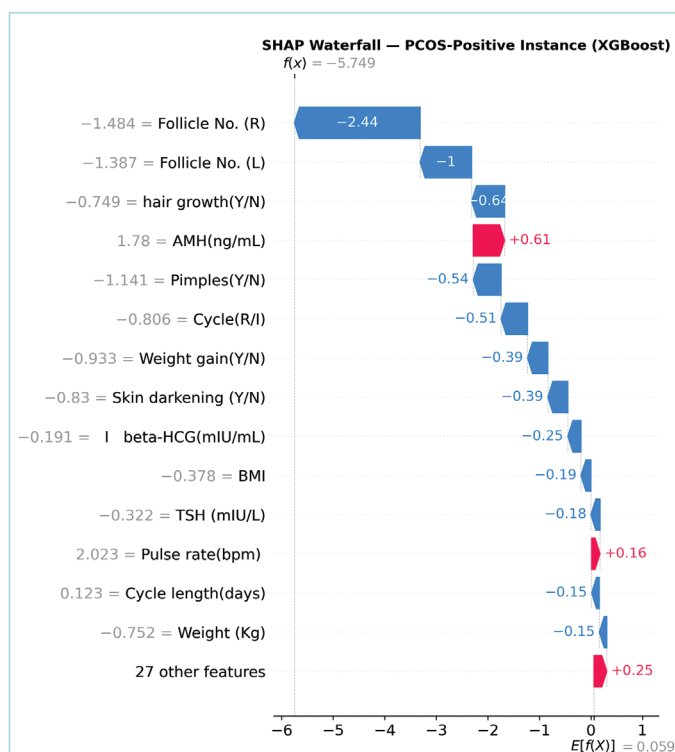


Figure 8. SHAP waterfall plot for a PCOS-positive test instance (XGBoost). Red bars indicate positive contributions toward PCOS; blue bars indicate negative contributions.

SHAP: SHapley Additive exPlanations, XGBoost: Extreme gradient boosting, PCOS: Polycystic ovary syndrome, AMH: Anti-Müllerian hormone, beta-HCG: Beta-human chorionic gonadotropin, BMI: Body mass index, TSH: Thyroid-stimulating hormone.

are marginally lower than those from a pre-fold SMOTE pipeline, confirming the correction eliminated a small but real inflation artefact, in alignment with best practices.^{17,18}

The SHAP analysis provided both global and local interpretability layers. Importantly, the alignment between SHAP-ranked predictors and components of the Rotterdam diagnostic criteria should be interpreted as internal consistency rather than independent discovery. The five most influential SHAP predictors - Follicle No. (R), hair growth (hirsutism), Follicle No. (L), weight gain, and cycle irregularity - are either direct components of or strongly associated with the Rotterdam criteria. This raises the possibility that the models are partially recapitulating the existing diagnostic rule. While this does not invalidate the models' discriminative utility, their primary contribution should be framed as the automated integration of established clinical features rather than as a novel biomarker discovery tool. The alignment between SHAP-ranked predictors and Rotterdam criteria components does not constitute independent discovery; rather, it confirms internal consistency between the model's learned feature weights and existing clinical knowledge, which is necessary but not sufficient for clinical trustworthiness.

AMH also appeared among the influential SHAP features, consistent with its emerging role as a PCOS biomarker and its relationship with symptom severity.^{22,23} Although included among the 41 input variables, the hormonal features FSH, LH, and their ratio were not among the top SHAP contributors. This finding likely reflects multicollinearity among FSH, LH, and other correlated hormonal variables (e.g., AMH, prolactin) within this dataset. Clinicians should, therefore, interpret the low SHAP ranking for FSH and LH with caution and in the context of established endocrine assessment.

The sensitivity analysis, excluding direct Rotterdam-criteria variables, yielded AUC-ROC values of 0.88 and 0.87 for XGBoost and LightGBM, respectively - a modest reduction that confirms the models retain discriminative capability based on non-criterion features (notably AMH, BMI, weight gain, hirsutism, and acne). As recently demonstrated by Ata et al.,²³ AMH shows significant associations with the severity of clinical PCOS symptoms, supporting its role as an informative non-criterion predictor.

The LIME explanations reinforced the SHAP global findings, with concordant identification of Follicle No. (R) and cycle irregularity as locally dominant features. This cross-method consistency between SHAP and LIME enhances the reliability of the interpretability analysis.²⁴

Calibration analysis indicated satisfactory calibration of predicted probabilities for both models (Brier scores 0.071 and 0.079), and DCA demonstrated positive net benefit compared with treat-all and treat-none strategies across clinically plausible threshold ranges. Despite these encouraging findings, formal prospective evaluation in real-world clinical workflows with human-in-the-loop validation is required before any recommendation for clinical deployment.

Study Limitations

Several limitations warrant acknowledgement. First, this study is based on a single, publicly available dataset compiled from ten hospitals in one geographic region (Kerala, India). External validation across ethnically and clinically diverse populations is required before generalisation can be assumed, given the well-documented variation in PCOS phenotypic expression across ethnicities.²⁵ Second, the models'

primary predictors are closely aligned with Rotterdam diagnostic criteria, raising the possibility of criterion-embedded feature dominance; clinicians should interpret these models as clinical integration tools rather than novel biomarker discovery systems. Specifically, three of the five most influential predictors-right and left follicle counts, and cycle irregularity-are themselves direct components of the Rotterdam criteria; therefore, a portion of the models' apparent performance necessarily reflects re-encoding of the diagnostic definition rather than an independent predictive signal. Third, while SMOTE effectively addresses class imbalance at the training stage, it generates synthetic rather than real patient data. Fourth, hyperparameter optimisation was performed via RandomizedSearchCV with nested cross-validation; Bayesian optimisation may offer further refinement. Fifth, the hold-out test set contains only 36 PCOS-positive cases, meaning that each misclassification shifts sensitivity estimates by approximately 2.8 percentage points. Sixth, both the hold-out evaluation and cross-validation constitute internal validation only; the absence of an independent external dataset is acknowledged as the study's most important limitation. Future directions include prospective multicentre validation in diverse populations, integration with deep learning approaches for imaging data, and development of a deployable clinical decision support interface incorporating real-time XAI explanations.

CONCLUSION

This study demonstrates that gradient boosting ensemble classifiers - XGBoost and LightGBM - can achieve clinically interpretable, highly accurate detection of PCOS from structured clinical and hormonal data. Both models achieved AUC-ROC values above 0.95 on the internal hold-out test set and above 0.97 in the corrected 5-fold cross-validation, outperforming all three baseline comparators. SHAP and LIME explanations consistently identified Follicle No. (R), hirsutism, Follicle No. (L), weight gain, and cycle irregularity as the five most influential predictors which is consistent with established PCOS diagnostic criteria and confirms the biological plausibility and interpretive transparency of the proposed pipeline. Because these dominant predictors are themselves embedded within the diagnostic criteria, the pipeline is best interpreted as automating the integration of established clinical features rather than an independent biomarker-discovery tool.

These findings provide proof-of-concept evidence for explainable gradient-boosting pipelines for PCOS decision support. The encouraging results from calibration and decision-curve analyses suggest potential clinical utility; however, prospective external validation in independent, ethnically diverse cohorts is necessary before recommending clinical deployment.

The proposed pipeline-combining SMOTE-augmented gradient boosting with multi-layer XAI - provides a replicable, extensible framework for PCOS decision support, adaptable to other gynaecological and endocrine classification tasks, such as endometriosis staging or gestational diabetes risk stratification.

MAIN POINTS

- Two gradient boosting classifiers, extreme gradient boosting and light gradient boosting machine, achieved high discriminative performance for polycystic ovary syndrome (PCOS) detection in internal validation (area under the curve-receiver operating characteristic >0.95 on hold-out testing and >0.97 in cross-

validation), outperforming logistic regression, random forest, and support vector machine baselines.

- Applying synthetic minority oversampling technique strictly within each cross-validation fold removed an optimistic performance-inflating artefact, underscoring the importance of leakage-free resampling when modelling imbalanced clinical data.
- SHapley Additive exPlanations and local interpretable model-agnostic explanations converged on follicle counts, hirsutism, weight gain, and cycle irregularity as the dominant predictors, providing transparent and clinically plausible interpretations of both global and individual model behaviour.
- Because several of these leading predictors are themselves components of the Rotterdam diagnostic criteria, the models are best understood as automating the integration of established clinical features rather than discovering new biomarkers.
- The pipeline should therefore be regarded as a proof-of-concept for explainable PCOS decision support; prospective external validation on independent, ethnically diverse cohorts is required before any clinical use.

ETHICS

Ethics Committee Approval: This study used a publicly available, fully de-identified dataset and involved no direct patient contact or intervention. Therefore, ethics committee approval was not required.

Informed Consent: Not applicable, as the study was based solely on anonymized, publicly available secondary data.

Footnotes

Authorship Contributions

Surgical and Medical Practices: Ü.K.D., Ş.Y., E.K., O.D., Concept: Ü.K.D., E.H.C., O.D., Design: Ü.K.D., Ş.Y., E.H.C., E.K., O.D., Data Collection and/or Processing: Ü.K.D., Ş.Y., E.H.C., E.K., Analysis and/or Interpretation: Ü.K.D., Ş.Y., E.H.C., E.K., O.D., Literature Search: Ü.K.D., Ş.Y., E.K., O.D., Writing: Ü.K.D.

DISCLOSURES

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

Declaration on the Use of Artificial Intelligence (AI): No artificial intelligence tools were used in the preparation of this manuscript.

REFERENCES

1. 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. *Hum Reprod.* 2016; 31(12): 2841-55.
2. Azziz R, Carmina E, Chen Z, Dunaif A, Laven JS, Legro RS, et al. Polycystic ovary syndrome. *Nat Rev Dis Primers.* 2016; 2: 16057.
3. Teede HJ, Misso ML, Costello MF, Dokras A, Laven J, Moran L, et al; International PCOS Network. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Hum Reprod.* 2018; 33(9): 1602-18.
4. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril.* 2004; 81(1): 19-25.
5. Teede HJ, Tay CT, Laven JJE, Dokras A, Moran LJ, Piltonen TT, et al. Recommendations from the 2023 International Evidence-based Guideline for the assessment and management of polycystic ovary syndrome. *J Clin Endocrinol Metab.* 2023; 108(10): 2447-69.
6. Sirmans SM, Pate KA. Epidemiology, diagnosis, and management of polycystic ovary syndrome. *Clin Epidemiol.* 2013; 6: 1-13.
7. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019; 25(1): 44-56.
8. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med.* 2016; 375(13): 1216-9.
9. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2016: 785-94.
10. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3149-57.
11. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001; 29(5): 1189-232.
12. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *arXiv.* 2017; 30.
13. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD.* 2016:1135-44.
14. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, González AB, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* 2020; 58: 82-115.
15. Kottarathil P. Polycystic Ovary Syndrome (PCOS) [dataset]. Kaggle; 2020. Available from: <https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002; 16: 321-57.
17. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics.* 2013; 14: 106.
18. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res.* 2017; 18: 1-5.
19. Denny A, Raj A, Ashok A, Ram CM, George R. i-HOPE: detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques. *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*. Kochi, India. 2019, pp. 673-8.
20. Mohi Uddin KM, Bhuiyan M.A, Rahman MM, Islam MM, Uddin MA. Early PCOS detection: a comparative analysis of traditional and ensemble machine learning models with advanced feature selection. *Engineering Reports.* 2025; 7(2): e70008.
21. Zigarelli A, Jia Z, Lee H. Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: observational study. *JMIR Form Res.* 2022; 6(3): e29967.
22. Tehrani FR, Solaymani-Dodaran M, Hedayat M, Azizi F. Is polycystic ovary syndrome an exception for reproductive aging? *Hum Reprod.* 2010; 25(7): 1775-81.
23. Ata C, Atlihan U, Avsar HA, Yavuz O, Erkilinc S, Bildaci TB. The relationship of antimüllerian hormone with severity of clinical symptoms in polycystic ovary syndrome. *BMC Womens Health.* 2025; 25(1): 432.

24. Molnar C. Interpretable machine learning: a guide for making black box models explainable. 2nd ed. 2022. Available at: <https://christophm.github.io/interpretable-ml-book/>
25. Teede H, Deeks A, Moran L. Polycystic ovary syndrome: a complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan. BMC Med. 2010; 8: 41.