Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Heterogeneous Stacking Ensemble Learning With Catboost Meta-Learner For Explainable and Statistically Validated PCOS Classification From Clinical Tabular Data

Swarnima Shrivastava[1]*, Varsha Thakur[2], Surendra Kumar Patel[3]

[1,3]Dept. of Information Technology, Govt. N.P.G. College of Science, Raipur, Chhattisgarh, India.
[2]Dept. of Computer Science, Govt. N.P.G. College of Science, Raipur, Chhattisgarh, India.
Email: swarnima.shrivastava@gmail.com

Abstract
Polycystic Ovary Syndrome (PCOS) is a hormonal disorder among reproductive-aged women, with approximately 70% of cases going undiagnosed due to the intricacy of existing diagnostic criteria. In this study, we performed a statistically validated comparative analysis of six classifiers, Random Forest(RF), AdaBoost, XGBoost, Support Vector Machine(SVM), Multi-Layer Perceptron(MLP), and a novel two-level stacking ensemble, to distinguish between PCOS and non-PCOS cases using the Kaggle PCOS dataset (n = 541). The proposed stacking ensemble combines five heterogeneous base learners through a hyperparameter-optimized CatBoost meta-learner trained on out-of-fold predictions, with SMOTE applied to address class imbalance. The implementation of stratified 10-fold cross-validation in conjunction with McNemar's statistical test represents a significant contribution of this research, an approach that has not been consistently documented in earlier studies. Our results indicate that the proposed stacking ensemble is the most effective model, demonstrating a test accuracy of 93.58%, precision of 96.77%, F1-score of 89.55%, and an AUC-ROC of 0.9680. Furthermore, cross-validation yielded a mean AUC-ROC of 0.9875 ± 0.0111. McNemar's test corroborated the statistically significant superiority of the ensemble relative to both Random Forest and AdaBoost. Additionally, SHAP enabled global interpretability of the optimized XGBoost model, while LIME provided patient-specific explanations for the stacking ensemble, identifying follicle count, hirsutism, weight gain, skin darkening, and menstrual irregularity as critical predictors, establishing a robust foundation for explainable clinical decision-support systems.

Keywords: polycystic ovary syndrome, stacking ensemble, machine learning, explainable artificial intelligence, McNemar's test, class imbalance.

1 Introduction

Polycystic ovary syndrome (PCOS) is identified as one of the most common endocrine and metabolic disorders impacting women for the duration of their reproductive years, affecting an estimated 5–10% of this population, with nearly 70% of cases remaining undiagnosed despite a rising global incidence(Rahman et al. 2024a; Uddin et al. 2025). Diagnosis remains inherently challenging, as it requires reconciling heterogeneous hormonal, clinical, and morphological criteria that overlap considerably with other endocrine disorders, complicating timely and consistent classification(Firdatunnisa et al. 2026a; Zad et al. 2024a). Beyond its reproductive implications, untreated PCOS is strongly connected with long-term metabolic and cardiovascular issues, together with insulin resistance, hypertension, and infertility (Faris et al. 2025). Conventional diagnostic pathways, which depend on costly hormonal assays and ultrasonographic assessment, are further constrained by inter-observer variability, limited dataset standardization, and inadequate multimodal integration(Li et al. 2025; Panjwani et al. 2025). Latest systematic reviews approve that artificial intelligence and machine learning demonstrate substantial diagnostic and predictive promise for PCOS, while also underscoring persistent gaps in generalizability, external validation, and standardized reporting(Ghaderzadeh et al. 2025a; Yao et al. 2025a). Building on this trajectory, recent frameworks employing feature selection, ensemble learning, and explainable AI have sought to jointly improve diagnostic accuracy and clinical interpretability in PCOS classification (Devi and Nivedha S 2026; Emara et al. 2025; Putri and Amriza 2026a).

Despite substantial progress, PCOS classification research using clinical tabular data continues to exhibit unresolved methodological gaps. Recent taxonomic and systematic-review analyses confirm that intelligent PCOS detection still lacks standardized feature taxonomies and a unified understanding of existing tools' comparative capabilities, with individual studies benchmarked under heterogeneous protocols rather than a controlled, paradigm-spanning comparison(Arora et al. 2025). Stacking-based efforts to date have also relied on homogeneous, same-family base learners — such as a two-tier Random Forest architecture — rather than a heterogeneous ensemble combined with a systematically hyperparameter-optimized meta-learner(G and P 2024). Explainability remains narrow in scope: existing frameworks typically report only global SHAP-based feature attribution or a SHAP-derived nomogram, without complementary local explanations or explicit validation against Rotterdam diagnostic criteria(Firdatunnisa et al. 2026a; Yao et al. 2025a). Class-imbalance correction is applied without explicit reporting of leakage-safe, fold-aware protocols(Uddin et al. 2025), ensemble comparisons are frequently presented without formal statistical significance testing(Panjwani et al. 2025), and recent systematic reviews of the field confirm that this lack of statistical validation remains largely unaddressed(Ghaderzadeh et al. 2025a).

Research Objectives and Scope

- This study first aims to establish a controlled, unified benchmark evaluating Random Forest, AdaBoost, XGBoost, SVM, MLP, and the proposed stacking ensemble under identical preprocessing, feature selection, and stratified 10-fold cross-validation for PCOS classification.
- Building on this benchmark, it aims to determine whether a heterogeneous stacking ensemble combining diverse base learners via out-of-fold meta-feature generation with a hyperparameter-optimized CatBoost meta-learner can outperform the individually benchmarked classifiers.
- The study further aims to evaluate the capacity of an integrated global (SHAP) and local (LIME) explainability framework to identify feature importances that are congruent with the Rotterdam diagnostic criteria.
- Additionally, it seeks to determine whether the performance enhancements achieved by the proposed stacking ensemble retain their robustness and statistical significance when assessed through leakage-free, fold-aware preprocessing techniques and McNemar-based significance testing.

Contributions

This study provides four original contributions.

- First, it establishes a controlled, unified benchmark evaluating six classifiers—Random Forest, AdaBoost, XGBoost, SVM, MLP, and the proposed stacking ensemble—under identical preprocessing, feature selection, fold-aware SMOTE balancing, and stratified 10-fold cross-validation, enabling a rigorous, reproducible comparison of diverse learning paradigms for PCOS prediction.
- Second, it introduces a novel stacking ensemble in which five algorithmically diverse base learners are combined via out-of-fold meta-feature generation with a hyperparameter-optimized CatBoost meta-learner, achieving a test-set AUC-ROC of 0.9680 and a 10-fold CV AUC of 0.9875 ± 0.0111.
- Third, it develops an integrated explainability framework combining SHAP-based global analysis on the tuned XGBoost classifier with LIME-based local explanations on the stacking ensemble, identifying ovarian follicle count, menstrual cycle characteristics, hyperandrogenic manifestations, and AMH-related indicators as influential predictors—findings that align closely with the Rotterdam diagnostic criteria.
- Finally, the framework ensures robust validation through leakage-free preprocessing, fold-aware SMOTE balancing, stratified 10-fold cross-validation, and McNemar-based significance testing, proposing a more rigorous assessment of classifier performance and predictive stability than is typically reported in existing PCOS machine learning studies.

The remainder of this paper is structured as follows: Section 2 reviews related literature work. Section 3 describes the Materials and Methods. Section 4 presents results. Section 5 represents a discussion of the results. Section 6 concludes.

2 Literature Review

Research into AI-assisted assessment of polycystic ovary syndrome (PCOS) has moved beyond simple comparison of classifier performance toward more integrated predictive systems—ones that bring together feature selection, ensemble modeling, explainability, and deployment within digital health tools. Research in this domain often demonstrates notable discriminative performance utilizing structured clinical data. However, the variability in methods adopted by researchers concerning preprocessing, management of class imbalance, validation procedures, and interpretation of results complicates direct comparisons of

findings across different studies. From the viewpoint of health technology, it is essential to note that high predictive accuracy, in isolation, does not necessarily equate to substantive clinical value. Equally important are reproducibility, transparency, robustness when applied to different populations, and the ability of these systems to assist instead of alternate for clinical decision-making. With this in mind, the existing literature is critically reviewed across four dimensions: machine-learning-based screening, ensemble-based prediction, explainable AI methods, and the state of validation and readiness for real-world translation.

2.1 Machine Learning for Technology-Assisted PCOS Screening

Early machine-learning research established the feasibility of identifying PCOS from structured demographic, hormonal, metabolic, reproductive, and ultrasound-derived variables. Nasim et al. (Nasim et al. 2022) combined chi-squared feature selection with extensive classifier benchmarking and reported near-perfect performance for Gaussian Naïve Bayes, whereas Danaei Mehr and Polat(Danaei Mehr and Polat 2022) obtained 98.89% accuracy and 100% sensitivity using embedded feature selection with an ensemble Random Forest. Rahman et al.(Rahman et al. 2024b) similarly demonstrated competitive performance for AdaBoost and Random Forest following Mutual-Information-based feature selection. Together, these findings specify that feature representation can influence PCOS classification as strongly as classifier choice. Nevertheless, exceptionally high performance obtained from small benchmark datasets should be interpreted cautiously when independent external evaluation is absent, because repeated optimization on a single cohort may overestimate transportability(Arora et al. 2025; Barrera et al. 2023).

Evidence from larger and clinically richer datasets illustrates the distinction between benchmark performance and generalizable health-technology value. Zad et al.(Zad et al. 2024b) evaluated logistic regression, SVM, gradient-boosted trees, and a multilayer-perceptron-based hormone score using an electronic health record cohort exceeding 30,000 patients, reporting out-of-sample AUC values between 0.80 and 0.85. Although lower than several results reported on the widely used 541-record Kaggle dataset, these estimates derive from a substantially larger clinical cohort. Reviews identify persistent limitations in the PCOS-AI evidence base, including small or homogeneous cohorts, inconsistent diagnostic definitions, variable feature-selection procedures, and insufficient external validation (Arora et al. 2025; Barrera et al. 2023; Ghaderzadeh et al. 2025b). Direct comparison of accuracy values across studies is therefore inappropriate unless dataset composition and evaluation protocols are considered simultaneously.

A parallel strand of research has moved from retrospective classification toward technology-assisted screening and patient support. Nimmala et al(Nimmala et al. 2024) integrated machine-learning prediction with an online support platform, while Prasher et al.(Prasher and Nelson 2025) emphasized non-invasive indicators and questionnaire information for accessible early screening. Multi-level frameworks have also contrasted clinically intensive prediction models with simplified screening-oriented architectures (Chiliveri et al. 2026). These studies are significant from a health-technology standpoint because they address accessibility and implementation rather than predictive accuracy alone. However, deployment-oriented prototypes do not themselves establish clinical effectiveness, and reduced-input models may sacrifice information contained in hormonal or imaging variables.

2.2 Ensemble Learning and Advanced Predictive Frameworks

Ensemble learning has rapidly been used to exploit complementary decision boundaries across classifiers. Alshakrani et al.(Alshakrani S, Hilal S, Zeki AM 2022) used five base learners and a gradient-boosting meta-learner, achieving 95.7% accuracy after PCA-based feature reduction. Emara et al. (Emara et al. 2025) integrated class balancing and BORUTA feature selection with heterogeneous base learners and an XGBoost meta-classifier, reporting 97% accuracy. Uddin et al.(Uddin et al. 2025) combined feature selection, SMOTE, and a soft-voting MLP–CatBoost ensemble, obtaining 96.88% accuracy, while Akter and Reno(Akter and Reno 2025a) reported 94.34% accuracy and 93.38% AUC using a clinically informed weighted ensemble. These findings propose that ensemble methods can improve predictive stability; however, their methodological contribution cannot be inferred from accuracy alone when preprocessing, resampling, and validation differ across studies.

Stacking is particularly relevant because it learns how to combine base-model outputs rather than applying fixed voting rules. SmartScanPCOS(G and P 2024) employed a hierarchical Random-Forest-based architecture and reported 99.31% accuracy, whereas Elmannai et al.(Elmannai et al. 2023) combined stacking with SMOTE-ENN, recursive feature elimination, and SHAP/LIME interpretation and reported 100% accuracy on a single train–test split. Such results demonstrate the capacity of optimized pipelines to fit structured PCOS data but also reinforce concerns regarding dataset-specific optimization without independent validation. From a methodological perspective, stacking requires careful construction of meta-level training data so that the meta-learner is not trained on predictions produced from the same samples used to fit the corresponding base models. In the present study, this requirement motivates the use of out-of-fold predictions for meta-feature generation.

The literature also shows a transition from simple ensemble aggregation toward integrated health-AI pipelines. Reka et al.(Reka et al. 2025) combined ensemble filtering, reinforcement-learning-enhanced optimization, and SHAP/LIME-based interpretation, while other studies incorporated web deployment, blockchain-based data management, or screening-oriented interfaces (Jaganathan and Natesan 2025; Nimmala et al. 2024; Uddin et al. 2025). These developments broaden the technological scope of PCOS prediction, but increasing architectural complexity may reduce reproducibility and complicate clinical implementation. Moreover, most high-performing ensemble studies remain dependent on retrospective benchmark datasets. A need therefore remains for controlled comparisons in which conventional learners, boosting methods, neural models, and heterogeneous stacking are evaluated under an identical experimental protocol.

2.3 Explainable AI and Clinical Interpretability

The growing use of complex ensemble and neural models has made interpretability a central requirement for AI-enabled health technologies. A systematic review of diagnostic and predictive AI applications in PCOS documented the growing use of explainability techniques, including SHAP, LIME, and Grad-CAM(Ghaderzadeh et al. 2025b). Khanna et al. (Khanna et al. 2023) employed multiple explanation techniques alongside a multi-stack learning framework, while Patil et al. (Patil et al. 2025) and Athapaththu and Malkanthi(Athapaththu and Malkanthi 2026) used SHAP to interpret feature-selection-driven models. Across several explainable PCOS classification studies, ovarian follicle characteristics, AMH, menstrual-cycle variables, weight-related characteristics, and manifestations associated with hyperandrogenism have been reported among influential predictors(Athapaththu and Malkanthi 2026; Firdatunnisa et al. 2026b; Patil et al. 2025; Putri and Amriza 2026b). Although the exact ranking varies by dataset, feature-selection strategy, and predictive model, these findings show recurring emphasis on clinically relevant reproductive, hormonal, metabolic, and symptom-related information.

Interpretability, however, should not be equated with clinical validity. Post-hoc explanations characterize model behaviour but do not establish causal relationships or guarantee reliable performance in a different population. Salih et al. (Salih et al. 2025) emphasized the complementary roles of global and local explanation methods in healthcare machine learning. Yao et al.(Yao et al. 2025b) advanced this translational perspective through multicentre external validation of a SHAP-interpreted XGBoost model and its translation into a nomogram. In contrast, many PCOS studies apply XAI to models evaluated only on a single retrospective dataset (Athapaththu and Malkanthi 2026; Firdatunnisa et al. 2026b; Khanna et al. 2023; Patil et al. 2025; Putri and Amriza 2026b). For health-technology research, clinical trust additionally depends on validation, calibration, reproducibility, and evidence of reliable performance in the intended setting.

Explainability is also frequently incorporated selectively. Some studies report only global feature importance, whereas others employ multiple explanation techniques without clearly distinguishing whether they explain the same model or different components of a pipeline. Global and local explanations serve different purposes: population-level attribution identifies variables that systematically influence predictions, while patient-specific explanations support examination of individual model decisions. Their complementary use can strengthen interpretability when the model being explained and the scope of each explanation are explicitly reported.

2.4 Validation and Translation of AI-Based PCOS Screening Technologies

The principal barrier separating high-performing PCOS classifiers from clinically credible health technologies is the limited strength of validation evidence. Much of the literature is concentrated on the Kaggle PCOS dataset, enabling reproducible benchmarking but also repeated optimization on a relatively small, geographically constrained cohort. Reviews consistently identify limited external validation, inconsistent validation procedures, and inadequate reporting of data-processing pipelines as major threats to generalizability (Arora et al. 2025; Barrera et al. 2023; Ghaderzadeh et al. 2025b). Yao et al.(Yao et al. 2025b) represents an important exception through multicentre internal and external evaluation, while Akter and Reno(Akter and Reno 2025b) reported evaluation across geographic cohorts. Nevertheless, prospective evaluation in routine clinical workflows remains uncommon.

Class-imbalance handling and statistical comparison constitute additional sources of methodological variability. Oversampling techniques such as SMOTE are widely used, but applying resampling before data partitioning or outside cross-validation folds can introduce information leakage and optimistic estimates. Likewise, performance improvements are often reported from differences in accuracy or AUC without formal paired statistical testing. For health-technology evaluation, leakage-aware preprocessing, out-of-fold model construction, stratified validation, and appropriate statistical comparison are therefore necessary to distinguish reproducible improvements from variation caused by a particular data split.

Translation also requires attention to the intended role of the technology. The evidence supports AI as a potential adjunct for early risk stratification and screening, particularly where access to specialist assessment is limited (Chiliveri et al. 2026; Nimmala et al. 2024; Prasher and Nelson 2025). Current PCOS

prediction systems, however, should not be interpreted as autonomous replacements for established clinical assessment. Before implementation, models require independent external and prospective validation, evaluation of subgroup performance and calibration, assessment of missing-data robustness, and usability testing within real clinical workflows.

2.5 Synthesis of Research Gaps and Study Positioning

The literature demonstrates substantial progress in machine-learning-based PCOS prediction, yet several interrelated methodological gaps remain. Reviews of the field highlight heterogeneity in datasets, preprocessing, model development, and validation, together with limited external validation(Arora et al. 2025; Barrera et al. 2023; Ghaderzadeh et al. 2025b). Within the representative experimental studies examined here, ensemble architectures, imbalance handling, explainability, and statistical comparison are also implemented inconsistently. Consequently, direct attribution of reported performance gains to a particular model architecture remains difficult. These observations motivate a controlled evaluation in which diverse classifiers and a stacking ensemble are assessed under a common experimental framework, while the scope of each explainability method and the limits of single-dataset validation are explicitly reported.

Table 1: Comparative methodological analysis of existing machine learning-based PCOS prediction studies and the proposed framework.

Study	Feature Selection	Imbalance Handling	Cross-Validation	Hyperparameter Tuning	Ensemble Learning	Stacking	XAI	Statistical Test
Uddin et al.(Uddin et al. 2026)	✓	✓	✓	✓	✓	✗	✓	✗
Rahman et al. (Rahman et al. 2024b)	✓	✗	✓	✗	✓	✗	✗	✗
Zad et al.(Zad et al. 2024b)	✓	✗	✗	✓	✗	✗	✗	✗
Yao et al.(Yao et al. 2025b)	✓	✗	✓	✓	✗	✗	✓	✗
Sahoo et al. (Sahoo et al. 2026)	✓	✓	✓	✓	✓	✗	✓	✗
Emara et al. (Emara et al. 2025)	✓	✓	✓	✓	✓	✓	✗	✗
Mahesswari & Maheswari (SmartScanPCOS) (G and P 2024)	✓	✗	✓	✓	✓	✓	✓	✗
Nasim et al.(Nasim et al. 2022)	✓	✗	✓	✗	✗	✗	✗	✗
Alshakrani et al. (Alshakrani S, Hilal S, Zeki AM 2022)	✓	✗	✓	✗	✓	✓	✗	✗
Danaei Mehr & Polat (Danaei Mehr and Polat 2022)	✓	✗	✓	✗	✓	✗	✗	✗
Elmannai et al. (Elmannai et al. 2023)	✓	✓	✗	✓	✓	✓	✓	✗

Study	Feature Selection	Imbalance Handling	Cross-Validation	Hyperparameter Tuning	Ensemble Learning	Stacking	XAI	Statistical Test
Akter & Reno (Akter and Reno 2025b)	✓	✓	✓	✓	✓	✗	✓	✗
Khanna et al. (Khanna et al. 2023)	✓	✓	✓	✓	✓	✓	✓	✗
Patil et al. (Patil et al. 2025)	✓	✗	✓	✓	✗	✗	✓	✗
Proposed Work	✓	✓	✓	✓	✓	✓	✓	✓

Against this background, the present study is positioned as a controlled methodological evaluation of PCOS classification from structured clinical data. It compares Random Forest, AdaBoost, XGBoost, SVM, MLP, and a heterogeneous stacking ensemble under a common framework; constructs stacking meta-features through out-of-fold predictions; uses a hyperparameter-optimized CatBoost meta-learner; and applies fold-aware imbalance handling. Interpretability is examined through global SHAP analysis of tuned XGBoost and local LIME explanations of the stacking ensemble, while McNemar's test assesses pairwise performance differences. The study therefore addresses methodological rigor, interpretability, and statistical validation within a single framework, while recognizing that independent multicentre validation remains necessary before clinical translation.

3. Materials and Methods

This section presents the complete experimental framework adopted in this study.

3.1 Dataset Description

The publicly available Kaggle PCOS dataset, originally compiled from tertiary care hospital records in Kerala, India using the Rotterdam diagnostic criteria, serves as the primary data source for this study (Kottarathil 2020). It comprises 541 patient records encompassing 41 multi-modal attributes — spanning hormonal, biochemical, metabolic, anthropometric, ultrasound, and lifestyle domains — with a binary class label indicating PCOS-positive (1) or PCOS-negative (0) (Agirsoy and Oehlschlaeger 2025). The dataset has been independently adopted as a benchmark in multiple PCOS classification studies(Nayem et al. 2025),(Jaganathan and Natesan 2025) enabling rigorous and reproducible comparison of the proposed framework against prior methods.

The class distribution is imbalanced, with approximately 34.4% PCOS-positive cases (n=177) and 65.6% PCOS-negative cases (n=364), a ratio of approximately 1:2. This imbalance is addressed in the preprocessing stage using SMOTE, as detailed in Section 3.3.

3.2. Workflow Overview

The overall workflow of the proposed framework is illustrated in Fig 1. The process begins with the acquisition of the Kaggle PCOS dataset, followed by data preprocessing, including feature scaling and class imbalance handling. To recognize the most informative predictors and reduce feature dimensionality, Mutual Information-based feature selection is then applied.

Subsequently, five diverse base learners—Random Forest, AdaBoost, XGBoost, Support Vector Machine, and Multi-Layer Perceptron—are employed within a stratified 10-fold cross-validation framework. The out-of-fold predictions generated by these base learners are combined to create a meta-feature matrix. This matrix subsequently serves as the input for a CatBoost meta-learner that has undergone hyperparameter-optimization. The final predictions from the stacking ensemble are evaluated using standard classification metrics and McNemar's statistical test. To enhance transparency and clinical interpretability, the SHAP analysis is utilized to evaluate the global feature importance of the optimized XGBoost model, whereas LIME is employed to produce individualized explanations for the suggested stacking ensemble.

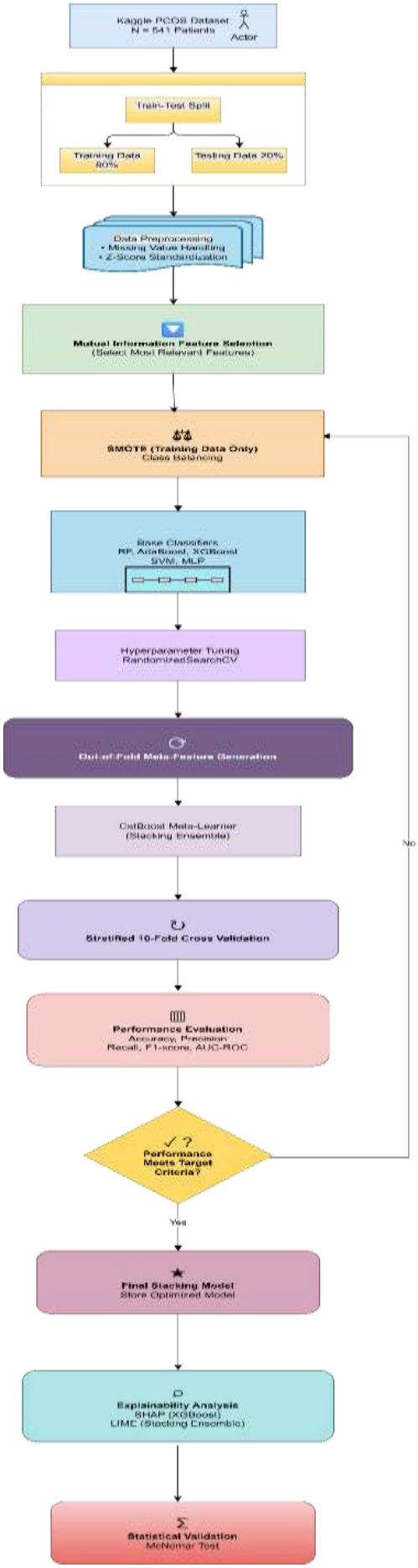
Kaggle PCOS Dataset
N = 541 Patients
Actor
Train-Test Split
Training Data
80%
Testing Data 20%
Data Preprocessing
• Missing Value Handling
• Z-Score Standardization
Mutual Information Feature Selection
(Select Most Relevant Features)
SMOTE (Training Data Only)
Class Balancing
Base Classifiers
RF, AdaBoost, XGBoost
SVM, MLP
Hyperparameter Tuning
RandomizedSearchCV
Out-of-Fold Meta-Feature Generation
CatBoost Meta-Learner
(Stacking Ensemble)
Stratified 10-Fold Cross Validation
Performance Evaluation
Accuracy, Precision
Recall, F1-score, AUC-ROC
Performance
Meets Target
Criteria?
No
Yes
Final Stacking Model
Store Optimized Model
Explainability Analysis
SHAP (XGBoost)
LIME (Stacking Ensemble)
Statistical Validation
McNemar Test

Fig. 1 Methodology flowchart of the proposed heterogeneous stacking ensemble framework for explainable and statistically validated PCOS classification using clinical tabular data

3.3 Novelty of the Proposed Approach

A detailed assessment of the existing literature of Section 2, reveals significant limitations in current PCOS detection approaches. Most existing studies employ single classifiers or simple voting ensembles without addressing the class imbalance problem within cross-validation folds, leading to data leakage. Furthermore, explainability is either absent or limited to a single method, and rigorous statistical validation through significance testing is rarely performed. To address these gaps, the proposed framework makes the following novel contributions: (i) a heterogeneous stacking ensemble of five diverse base learners with a CatBoost meta-learner, (ii) SMOTE applied strictly within each training fold to prevent leakage, (iii) dual XAI using SHAP for global and LIME for local explanation, and (iv) McNemar's test for statistical significance validation.

Table 2 presents a structured summary of the methodological contributions of the proposed framework, each directly addressing a limitation identified in the existing PCOS detection literature reviewed in Section 2.

Table 2: Summary of methodological contributions of the proposed heterogeneous stacking ensemble framework for PCOS detection.

Contribution	Detail
Ensemble architecture	Heterogeneous stacking — 5 base learners + CatBoost meta-learner
Imbalance handling	SMOTE inside each training fold — leakage-free
Explainability	SHAP (global feature ranking) + LIME (local per-sample)
Validation	Stratified 10-fold CV + McNemar's statistical significance test
Feature selection	Mutual information — applied before SMOTE

3.4 Train-Test Partitioning

Before any preprocessing transformations were applied, the complete dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ represents the feature vector of the i-th observation and $y_i \in \{0,1\}$ indicates its class label (0: PCOS-negative; 1: PCOS-positive), was divided into a training set $\mathcal{D}_{\text{train}}$ and a held-out test set $\mathcal{D}_{\text{test}}$ utilizing a stratified split with a ratio of 80:20 (Kaufman et al. 2012).

$$|\mathcal{D}_{\text{train}}| = \lfloor 0.8\, N \rfloor, \qquad |\mathcal{D}_{\text{test}}| = N - |\mathcal{D}_{\text{train}}|. \tag{1}$$

Stratification ensures that the empirical class distribution P^(y) is preserved across both subsets, thereby precluding distributional shift between training and evaluation partitions. All preprocessing estimators (imputer, scaler, feature selector) were fitted exclusively on $\mathcal{D}_{\text{train}}$ and subsequently applied as transformations to $\mathcal{D}_{\text{test}}$ strictly precluding any form of data leakage(Kaufman et al. 2012) .

3.5 Data Preprocessing

3.5.1 Missing Value Imputation

Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ represent the design matrix comprising n training samples and d features. In the case of missing entries in column j, these are substituted with the median of that column calculated from the available training set data.

$$\hat{x}_{ij} = \begin{cases} median\big(\{x_{kj} : x_{kj} \neq \text{NaN},\ k \in \mathcal{D}_{\text{train}}\}\big) & \text{if } x_{ij} = \text{NaN and } j \in \mathcal{F}_{\text{cont}}, \\ mode\big(\{x_{kj} : x_{kj} \neq \text{NaN},\ k \in \mathcal{D}_{\text{train}}\}\big) & \text{if } x_{ij} = \text{NaN and } j \in \mathcal{F}_{\text{cat}}, \end{cases} \tag{2}$$

where $\mathcal{F}_{\text{cont}}$ and $\mathcal{F}_{\text{cat}}$ denote the index sets of continuous and categorical features, respectively. The median imputation strategy is preferred over mean imputation for its robustness to skewed distributions and outliers, which are common in hormonal biomarker data(Little and Rubin 2019).

3.5.2 Feature Standardisation

To place all continuous features on a commensurate scale and to satisfy the distributional assumptions of gradient-based and kernel-based classifiers, z-score standardisation was applied to each continuous feature $j \in \mathcal{F}_{\text{cont}}$:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{3}$$

where

$$\mu_j = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} x_{ij}\,, \qquad \sigma_j = \sqrt{\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} \big(x_{ij} - \mu_j\big)^2} \tag{4}$$

Both μ_j and σ_j are estimated from $\mathcal{D}_{\text{train}}$ alone and applied identically to $\mathcal{D}_{\text{test}}$, ensuring that no test-set statistics contaminate the normalisation process(Bishop and Nasrabadi 2006).

3.6 Feature Selection

To minimize feature redundancy and enhance the interpretability of the model, a filter-based feature selection strategy was implemented. Mutual Information (MI) (Peng et al. 2005) was utilized to assess the dependence between each feature and the target variable. In contrast to linear correlation measures, MI effectively captures both linear and non-linear relationships by evaluating the extent of information a feature conveys regarding the class label. Features were ranked according to their mutual information scores using the SelectKBest method with the mutual_info_classif criterion, and the top-ranked features were retained for subsequent model development and evaluation.

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p\,(x,y)\; log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{5}$$

where p(x,y) denotes the joint probability distribution of feature X and class label Y, and p(x) a) denote their marginal probability distributions. A higher mutual information value indicates a stronger dependency between the feature and the target variable, implying greater relevance for classification(Peng et al. 2005).

3.7 Synthetic Minority Over-Sampling (SMOTE)

Class imbalance, wherein the number of PCOS-negative instances substantially exceeds that of PCOS positive instances, biases trained classifiers toward the majority class and yields deceptively high overall accuracy at the cost of poor minority-class recall(He and Garcia 2009). The Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002) was applied exclusively to $\mathcal{D}_{\text{train}}$ to produce a balanced training set.

For a minority-class instance $\boldsymbol{x}_i$, SMOTE selects one of its k-nearest neighbours $\boldsymbol{x}_{i_{(r)}}$ (with k = 5 by default) uniformly at random and generates a synthetic sample $\tilde{\boldsymbol{x}}$ by linear interpolation along the feature-space segment connecting the two:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}_i + \lambda\left(\boldsymbol{x}_{i_{(r)}} - \boldsymbol{x}_i\right), \qquad \lambda \sim \mathcal{U}(0,1), \tag{6}$$

where $\mathcal{U}(0,1)$ denotes the standard uniform distribution. This oversampling is repeated until the desired inter-class balance ratio is achieved. The test set $\mathcal{D}_{\text{test}}$ is never oversampled, thereby preserving the natural class distribution for evaluation(Blagus and Lusa 2013).

3.8 Baseline Classifiers

Five individual classifiers spanning distinct algorithmic families were implemented and evaluated as baselines.

3.8.1 Random Forest (RF)

Random Forest (Breiman 2001) is a bagging-based ensemble of T classification trees $\{h_t(\boldsymbol{x})\}_{t=1}^{T}$, each trained on a

bootstrap sample $\mathcal{D}_t \sim \mathcal{D}_{\text{train}}$ with replacement, and using a random subset of $m = \lfloor\sqrt{d}\rfloor$ features at each split. The final class prediction is determined by majority vote:

$$\hat{y}_{\text{RF}}(\boldsymbol{x}) = \underset{c\in\{0,1\}}{arg\,max} \sum_{t=1}^{T} \mathbf{1}\left[h_t(\boldsymbol{x}) = c\right] \tag{7}$$

The posterior class probability estimate is:

$$\hat{P}_{\text{RF}}(y = c \mid \boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[h_t(\boldsymbol{x}) = c\right] \tag{8}$$

Feature randomisation at each node decorrelates individual trees, reducing ensemble variance without substantially increasing bias(Hastie et al. 2009).

3.8.2 AdaBoost

AdaBoost (Adaptive Boosting) (Freund and Schapire 1997)constructs an additive ensemble of M weak learners $\{g_m(\boldsymbol{x})\}_{m=1}^{M}$ (shallow decision trees, typically stumps) through a sequential re-weighting scheme. At iteration m, the sample weight distribution $\mathcal{W}^{(m)} = \{w_i^{(m)}\}_{i=1}^{n}$ emphasises previously misclassified instances:

$$\epsilon_m = \sum_{i=1}^{n} w_i^{(m)}\,\mathbf{1}[g_m(\boldsymbol{x}_i) \neq y_i], \qquad \alpha_m = \frac{1}{2} ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right), \tag{9}$$

$$w_i^{(m+1)} = \frac{w_i^{(m)} exp(-\alpha_m\, y_i\, g_m(\boldsymbol{x}_i))}{Z_m}, \tag{10}$$

where Z_m is a normalisation constant. The final strong classifier is the weighted sign vote:

$$\hat{y}_{\text{AdaBoost}}(\boldsymbol{x}) = sign\big(\sum_{m=1}^{M} \alpha_m \; g_m(\boldsymbol{x})\big). \tag{11}$$

3.8.3 XGBoost (Extreme Gradient Boosting)

XGBoost (Chen and Guestrin 2016) is a regularised gradient boosting framework that builds an additive ensemble of K regression trees $\{f_k\}_{k=1}^{K}$ by minimising a second-order Taylor-expanded objective with explicit regularisation:

$$\mathcal{L}^{(k)} = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(k-1)} + f_k(\boldsymbol{x}_i)\right) + \Omega(f_k), \tag{12}$$

In this context, $\ell(\cdot)$ represents a differentiable loss function, specifically the binary cross-entropy utilized for classification tasks, while the regularisation term $\Omega(f_k)$ serves to penalise the complexity of the model.

$$\Omega(f_k) = \gamma\, T_k + \frac{1}{2}\lambda \parallel \boldsymbol{w}_k \parallel^2, \tag{13}$$

With T_k denoting the number of leaves, $\boldsymbol{w}_k$ the leaf weight vector, γ the minimum loss reduction for a split, and λ the L2 penalty coefficient. Using a second-order Taylor expansion of the loss, the optimal leaf weight for leaf j is:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}, \tag{14}$$

where $g_i = \partial_{\hat{y}}\ell(y_i, \hat{y}_i)$ and $h_i = \partial_{\hat{y}}^2\ell(y_i, \hat{y}_i)$ are the first- and second-order gradient statistics, respectively. The optimal split gain for a candidate split is:

$$\mathcal{G}_{\text{split}} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{15}$$

3.8.4 Support Vector Machine (SVM)

The Support Vector Machine (Cortes and Vapnik 1995) identifies the hyperplane with the highest margin that separates different classes within a kernel-induced feature space $\mathcal{H}$. This objective is articulated in the soft-margin primal formulation:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \; \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + C \sum_{i=1}^{n} \xi_i \quad \text{subject to} \quad y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{16}$$

where $C > 0$ is the regularisation parameter and $\phi: \mathbb{R}^d \to \mathcal{H}$ is the feature map. The dual formulation gives the decision function:

$$f(\boldsymbol{x}) = sign\big(\sum_{i=1}^{n_{\text{sv}}} \alpha_i \; y_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b\big), \tag{17}$$

where $\alpha_i \geq 0$ are Lagrange multipliers and $\kappa(\cdot,\cdot)$ is the Radial Basis Function (RBF) kernel:

$$\kappa\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) = exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right), \tag{18}$$

with bandwidth parameter σ (equivalently, $\gamma = 1/2\sigma^2$) controlling the locality of the kernel.

3.8.5 Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron (Goodfellow et al. 2016) is a comprehensive feedforward neural network characterised by a fully-connected architecture, consisting of an input layer, L hidden layers, and concluding with a sigmoid function in the output layer. The forward pass for hidden layer ℓ is:

$$\boldsymbol{a}^{(\ell)} = \sigma_{\text{act}}\big(\boldsymbol{W}^{(\ell)}\, \boldsymbol{a}^{(\ell-1)} + \boldsymbol{b}^{(\ell)}\big), \tag{19}$$

where $\boldsymbol{W}^{(\ell)}$ and $\boldsymbol{b}^{(\ell)}$ are the weight matrix and bias vector, and σ_{act} denotes the ReLU activation:

$$\sigma_{\text{act}}(z) = max(0, z). \tag{20}$$

The output layer applies the logistic sigmoid to produce a class-posterior probability:

$$\hat{P}(y = 1 \mid \boldsymbol{x}) = \sigma\big(\boldsymbol{w}^{(L+1)\top}\boldsymbol{a}^{(L)} + b^{(L+1)}\big), \qquad \sigma(z) = \frac{1}{1+e^{-z}}. \tag{21}$$

Parameters are optimised via back-propagation using the Adam optimizer(Khan et al. 2020) to minimise the binary cross-entropy loss:

$$\mathcal{L}_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\big[y_i \, log\, \hat{P}\,(y_i = 1 \mid \boldsymbol{x}_i) + (1 - y_i)\, log\, \hat{P}\,(y_i = 0 \mid \boldsymbol{x}_i)\big] \tag{22}$$

3.9 Hyperparameter Optimization

Hyperparameter selection has a crucial role in determining model generalization performance. In this study, hyperparameter optimization was undertaken through the application of Randomized Search in conjunction with stratified cross-validation. This methodology diverges from the more traditional exhaustive grid search by sampling parameter combinations from specified distributions, thus offering a computationally efficient means of navigating high-dimensional hyperparameter spaces and pinpointing near-optimal parameter configurations(Bergstra and Bengio 2012; Pedregosa et al. 2011). The objective function was articulated as the mean cross-validated AUC-ROC score calculated on the training dataset, and the selection of the optimal parameter configuration was based on achieving the highest validation performance.

The goal of the optimization process is articulated as follows:

$$\theta^* = arg \max_{\theta \in \Theta} \mathcal{F}(\theta) \tag{23}$$

where:

Θ= hyperparameter search space

θ = candidate hyperparameter configuration

$\mathcal{F}(\theta)$ = mean cross-validated AUC-ROC score

For the XGBoost classifier, the exploration of hyperparameters encompassed the number of estimators (n_estimators), maximum tree depth (max_depth), learning rate (learning_rate), subsampling ratio (subsample), column sampling ratio (colsample_bytree), and regularization parameters (Chen and Guestrin 2016). In parallel, the optimization of the CatBoost meta-learner was conducted concerning the number of boosting iterations (iterations), learning rate (learning_rate), tree depth (depth), L2 regularization coefficient (l2_leaf_reg), and border count (border_count) (Prokhorenkova et al. 2018).

3.10 Proposed Stacking Ensemble

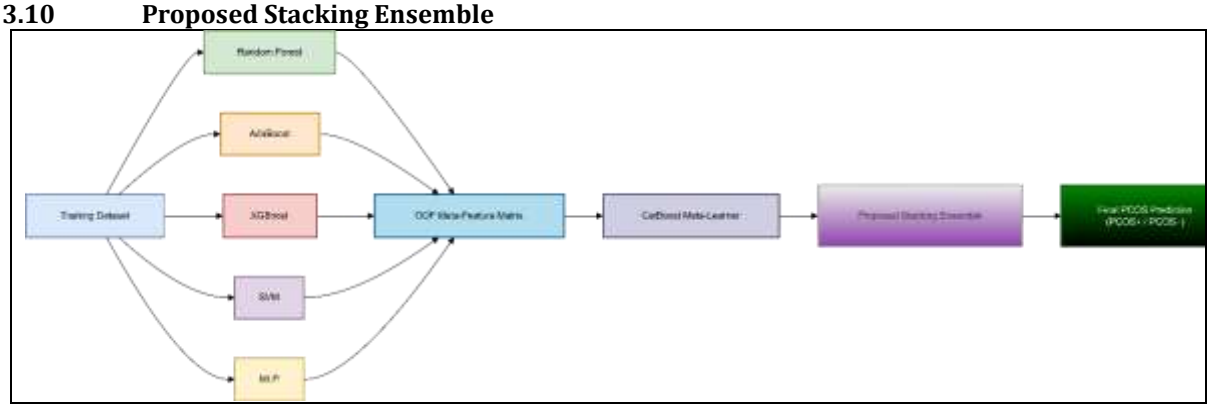

Fig. 2 Architecture of the proposed heterogeneous stacking ensemble. Probabilistic outputs generated by five base learners are combined into an out-of-fold (OOF) meta-feature matrix, which is subsequently used to train the CatBoost meta-learner for final PCOS classification

3.10.1 Stacking Generalization: Theoretical Basis

Stacking Generalization (SG) (Breiman 2001; Wolpert 1992) is a meta-learning paradigm in which a set of diverse base learners at Level-0 generate predictions that form the input feature matrix for a Level-1 meta-learner. The rationale is that a learnable combination of heterogeneous models can capture complementary inductive biases and reduce both bias and variance relative to any constituent model(LeBlanc and Tibshirani 1996).

3.10.2 Level-0: Base Learner Layer

Let $\mathcal{B} = \{RF, \text{AdaBoost}, \text{XGBoost}, SVM, MLP\}$ denote the set of $B = 5$ base learners. For each learner $b \in \mathcal{B}$, let $\phi_b: \mathbb{R}^d \to [0,1]$ denote its learned class-probability mapping after training on $\mathcal{D}_{\text{train}}$. The meta-feature representation of an instance $\boldsymbol{x}$ is formed by concatenating the positive-class probability estimates from all base learners:

$$\tilde{\boldsymbol{x}} = [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \dots, \phi_B(\boldsymbol{x})]^T \in [0,1]^B \tag{24}$$

3.10.3 Out-of-Fold Meta-Feature Generation

To mitigate overfitting of the meta-learner to training-set predictions and thereby preclude data leakage in the construction of the meta-feature matrix, out-of-fold (OOF) prediction with $K = 10$ folds are employed(Wolpert 1992). Let $\{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_K\}$ be a stratified K-fold partition of $\mathcal{D}_{\text{train}}$. For each fold k and each base learner b:

1. Train b on $\mathcal{D}_{\text{train}} \backslash \mathcal{F}_k$.
2. Predict on the held-out fold: $\hat{p}_{ib} = \phi_b^{(-k)}(\boldsymbol{x}_i)$ for $i \in \mathcal{F}_k$.

After all K folds, the complete OOF matrix $\widetilde{\boldsymbol{X}}_{\text{train}} \in \mathbb{R}^{|\mathcal{D}_{\text{train}}| \times B}$ is assembled, with each row $\widetilde{\boldsymbol{x}}_i = [\hat{p}_{i1}, \ldots, \hat{p}_{iB}]^{\top}$. For the test set, the prediction of the base learner b is computed as the mean of the predictions obtained from all K fold-specific models:

$$\hat{p}_{ib}^{\text{test}} = \frac{1}{K}\sum_{k=1}^{K} \phi_b^{(-k)}(\boldsymbol{x}_i), \quad i \in \mathcal{D}_{\text{test}} \tag{25}$$

3.10.4 Level-1: CatBoost Meta-Learner

The Level-1 meta-learner employs a hyperparameter-optimized CatBoost classifier. The selection of CatBoost is attributed to its robust gradient boosting framework, the implementation of an ordered boosting mechanism, and its notable generalization performance on structured tabular datasets. The ordered boosting strategy reduces prediction shift by constructing target statistics using only information accessible from preceding observations in a random permutation of the training data, thereby mitigating target leakage and overfitting(Prokhorenkova et al. 2018).

Let,

$$\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_B(\boldsymbol{x})]^T \tag{26}$$

denote the vector of meta-features generated by the (B) base learners. The CatBoost meta-learner is trained on the meta-feature matrix $\left(\widetilde{\boldsymbol{X}}_{train}\right)$ to learn an aggregation function

$$\hat{y} = f_{CB}(\widetilde{\boldsymbol{x}}) \tag{27}$$

where $f_{CB}(\cdot)$ denotes the trained CatBoost classifier and $\hat{y}$ represents the final predicted probability of PCOS. Through this process, complementary information from the diverse base learners is integrated into a single robust prediction.

3.10.5. Pseudocode of Proposed Stacking Ensemble

Algorithm 1: Proposed Heterogeneous Stacking Ensemble with CatBoost Meta-Learner

Input: Dataset D, Base learners B = {RF, AdaBoost, XGBoost, SVM, MLP}, K=10 folds
Output: Final PCOS prediction ŷ
Split D into D_train (80%) and D_test (20%) with stratification
Apply missing value imputation and z-score standardisation on D_train
Apply MI-based feature selection; retain top-k features
For each fold k = 1 to K:
Apply SMOTE on D_train \ F_k only
Train each b ∈ B on augmented D_train \ F_k
Generate OOF predictions p_i^b for i ∈ F_k
Assemble meta-feature matrix X_train from all OOF predictions
Train optimised CatBoost on X_train
Generate test meta-features by averaging K fold predictions
Predict ŷ = f_CB(X_test)
Evaluate using metrics; apply SHAP & LIME; run McNemar's test

3.11 Stratified K-Fold Cross-Validation

Model selection and unbiased performance estimation were conducted via stratified 10-fold cross-validation (10-FCV)(Kohavi and others 1995; Stone 1974). The training set is partitioned into ten mutually exclusive, approximately equal-sized stratified folds $\{\mathcal{F}_k\}_{k=1}^{10}$. In each iteration k, the model is trained on $\mathcal{D}_{\text{train}}^{(-k)} = \mathcal{D}_{\text{train}} \backslash \mathcal{F}_k$ and evaluated on $\mathcal{F}_k$. The generalized performance estimate for a metric $\mathcal{M}$ is:

$$\overline{\mathcal{M}} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{M}_k, \qquad s_{\mathcal{M}} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K} (\mathcal{M}_k - \overline{\mathcal{M}})^2}, \tag{28}$$

where $\overline{\mathcal{M}}$ and $s_{\mathcal{M}}$ denote the cross-validated mean and standard deviation, respectively. This procedure makes sure that every sample in $\mathcal{D}_{\text{train}}$ participates exactly once as a validation instance, providing a low-variance, approximately unbiased estimator of test performance(Arlot and Celisse 2010).

3.12 Performance Evaluation Metrics

The evaluation of classification performance was conducted through a detailed analysis employing a variety of metrics derived from the confusion matrix C. In this context, the matrix entries are defined as follows: TP represents true positives, FP signifies false positives, TN indicates true negatives, and FN corresponds to false negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \quad (29)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (30)$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}, \quad (31)$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad (32)$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (33)$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (34)$$

The Matthews Correlation Coefficient (MCC) (Matthews 1975) serves as a particularly valuable metric in scenarios characterised by class imbalance, as it comprehensively considers all four quadrants of the confusion matrix. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) (Hanley and McNeil 1982) was calculated as a threshold-independent indicator of the model's discriminatory efficacy.

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(t)\right) dt, \quad (35)$$

where TPR = Recall and FPR $= FP/(FP + TN)$ are evaluated at varying classification thresholds.

3.13 Explainable Artificial Intelligence

3.13.1 SHAP (SHapley Additive exPlanations)

To enhance transparency and ensure clinical interpretability, the SHAP method (Lundberg and Lee 2017) was applied to the tuned XGBoost classifier to quantify global feature importance and examine the direction and magnitude of feature contributions. SHAP grounds feature attribution in Shapley values from cooperative game theory (Shapley and others 1953): for a prediction f($\mathbf{x}$), the SHAP value $\phi_j(\boldsymbol{x})$feature j is the weighted average marginal contribution of feature j across all possible feature coalitions $S \subseteq \mathcal{F}\backslash\{j\}$:

$$\phi_j(x) = \sum_{S \subseteq F\backslash\{j\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \left[f_{S\cup\{j\}}\left(x_{S\cup\{j\}}\right) - f_S(x_S)\right] \quad (36)$$

where f_S denotes the model restricted to feature subset S. SHAP values satisfy the three axioms of efficiency, symmetry, and monotonicity, and they decompose any prediction additively:

$$f(\boldsymbol{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j\,(\boldsymbol{x}), \quad (37)$$

where $\phi_0 = \mathbb{E}[f(\boldsymbol{X})]$ is the baseline (expected model output). Global feature importance is assessed by the mean absolute SHAP value:

$$\bar{\phi}_j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left|\phi_j(\boldsymbol{x}_i)\right| \quad (38)$$

3.13.2 LIME (Local Interpretable Model-Agnostic Explanations)

LIME (Ribeiro et al. 2016) provides complementary local explanations by approximating the complex model f in the neighbourhood of an instance $\mathbf{x}_0$ with an interpretable surrogate model $g \in \mathcal{G}$ (linear model). LIME solves:

$$g^* = \underset{g \in \mathcal{G}}{arg\,min}\ \mathcal{L}\left(f,\, g,\, \pi_{x_0}\right) + \Omega(g), \quad (39)$$

where $\mathcal{L}\left(f, g, \pi_{x_0}\right) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \pi_{x_0}(\boldsymbol{z})(f(\boldsymbol{z}) - g(\boldsymbol{z}))^2$ is the local fidelity loss, $\pi_{x_0}(\boldsymbol{z}) = exp(-D(\boldsymbol{x}_0, \boldsymbol{z})^2/\delta^2)$ is an exponential proximity kernel weighting perturbed samples $\mathbf{z}$ by their distance $D(\cdot,\cdot)$ from the query instance, Z is the set of perturbed samples drawn from an interpretable representation, and Ω(g) penalises model complexity. The resulting linear coefficients of g^* quantify the directional feature contributions to the specific prediction f($\mathbf{x}_0$). SHAP and LIME collectively offer a robust framework for interpretability, combining both population-level (global) insights and instance-level (local) explanations. This dual approach enhances clinical trust and aids in the identification of biomarkers that are diagnostically significant.

3.14 Statistical Significance Analysis

To assess the statistical significance of the performance differences observed between classifiers, McNemar's test (McNemar 1947) was utilized. McNemar's test is a nonparametric statistical procedure for paired classification outcomes and is widely used to compare the predictive performance of two classifiers

evaluated on the same set of test instances. The null hypothesis ((H_0)) assumes that both classifiers exhibit identical error rates, whereas the alternative hypothesis ((H_1)) assumes a significant difference in their predictive behaviour.

Let n_{01} represent the number of instances that are accurately classified by classifier A yet misclassified by classifier B. Conversely, let n_{10} signify the number of instances that are accurately identified by classifier B but misclassified by classifier A. The statistic for the McNemar test is calculated as follows:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad (40)$$

The McNemar test statistic was calculated using the uncorrected asymptotic chi-square formulation, consistent with the values reported in Table 10.

Under the null hypothesis, it is expected that the test statistic will converge towards a chi-square distribution with one degree of freedom. The assessment of statistical significance was conducted at an alpha level of 0.05. A p-value that is less than 0.05 indicates a statistically significant difference in performance between the two classifiers, while a p-value that exceeds this threshold implies that the observed differences may be ascribed to random variation. To assess the statistical significance of the noted improvements in predictive performance, McNemar's test was employed for pairwise comparisons between the proposed stacking ensemble and each of the baseline classifiers.

3.15 Implementation Environment

All experiments are conducted using Python version 3.11 on a system equipped with an NVIDIA GPU utilizing CUDA version 12.2. A comprehensive overview of the software and hardware environment is presented in Table 3 below.

Table 3: Software and hardware environment

Component	Specification
Language	Python 3.11
Deep Learning framework	TensorFlow 2.15 / Keras
Machine Learning library	scikit-learn 1.4.2
Boosting Algorithm	XGBoost 2.0.3, CatBoost
Imbalance Handling	Imbalanced-learn 0.12
SHAP Library	shap 0.45.1
LIME Library	lime 0.2.0.1
Data Handling	pandas 2.2, NumPy
Visualisation	matplotlib 3.8, seaborn 0.13
OS	Ubuntu 22.04 LTS
CPU	Intel Core Ultra 5
GPU	NVIDIA RTX 3060
RAM	16GB

Reproducibility in this study is ensured by consistently setting the global random seed to 42 across all libraries employed, including NumPy, TensorFlow, scikit-learn, and XGBoost.

The following assumptions govern the experimental framework. Each patient is treated as a binary case (PCOS-positive or PCOS-negative), and multi-phenotype classification is outside the range of this study. The dataset is assumed to comprise independent patient records, with each observation representing a unique individual. All selected clinical features are assumed to be available at inference time. SMOTE (Chawla et al. 2002) is applied exclusively within the training folds to prevent information leakage. AUC-ROC is reported as a threshold-independent metric, whereas all other classification metrics are computed using a decision threshold of 0.5. For explainability analysis, SHAP was applied to the tuned XGBoost model to quantify global feature importance, while LIME was employed to provide local explanations for predictions generated by the proposed stacking ensemble. For McNemar's test (McNemar 1947), a non-significant result is interpreted as insufficient evidence to conclude a difference in classifier performance rather than evidence of equivalence.

4 Results and Discussion

This section exhibits the experimental results of all six evaluated classifiers on the Kaggle PCOS dataset: Random Forest (RF), AdaBoost, XGBoost, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and the proposed Stacking Ensemble with CatBoost meta-learner. Results include holdout test set performance,

confusion matrix analysis, ROC curve analysis, 10-fold cross-validation, SHAP global feature importance, LIME local explanations, and McNemar's statistical significance test. XAI analysis results from SHAP and LIME are presented in Section 4.6.

4.1 Individual Model Performance

Table 4 represents a comprehensive overview of the classification performance metrics for six models under evaluation: Random Forest (RF), AdaBoost, XGBoost, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and the proposed Stacking ensemble utilizing CatBoost as the meta-learner. The performance is assessed across five metrics: accuracy, precision, recall, F1-score, and AUC-ROC. Results are reported on the Kaggle PCOS dataset (N = 541). The proposed heterogeneous stacking ensemble achieved the best overall performance, attaining the highest accuracy (93.57%), precision (96.77%), F1-score (89.55%), and AUC-ROC (0.9680), although XGBoost and MLP achieved higher recall (86.11% versus 83.33%). These results demonstrate the overall predictive advantage of integrating diverse base learners through a CatBoost meta-learner while also highlighting the recall trade-off. The baseline models exhibited varying levels of predictive performance.

Table 4: Classification performance of all evaluated models on the Kaggle PCOS test set (best values in bold)

Model	Category	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
Stacking Ensemble	Ensemble	**[93.57]**	**[96.77]**	[83.33]	**[89.55]**	**[0.9680]**
XGBoost	Ensemble	[92.66]	[91.17]	**[86.11]**	[88.57]	[0.9585]
Random Forest	Ensemble	[90.82]	[88.23]	[83.33]	[85.71]	[0.9429]
SVM	Machine Learning	[92.66]	[93.75]	[83.33]	[88.23]	[0.9600]
AdaBoost	Ensemble	[89.90]	[85.71]	[83.33]	[84.50]	[0.9453]
MLP	Deep Learning	[90.82]	[86.11]	**[86.11]**	[86.11]	[0.9646]

All results in Table 4 are reported on the untouched 20% holdout test set. Stratified 10-fold cross-validation was conducted separately for model validation, with SMOTE applied only within the corresponding training folds.

4.2 Stratified 10-Fold Cross-Validation Results

Table 5 reports mean accuracy and standard deviation across 10 stratified folds for the proposed stacking ensemble. The framework achieved a mean accuracy of 94.22 ± 4.01%, precision of 95.43 ± 3.88%, recall of 93.00 ± 6.21%, F1-score of 94.09 ± 4.18%, and AUC-ROC of 0.9875 ± 0.0111, where the consistently low standard deviations across all metrics validate the stability and generalization capability of the proposed heterogeneous stacking architecture.

Table 5: 10-fold cross-validation results for the proposed stacking ensemble

Metric	Mean	Std Dev	Mean ± Std
Accuracy	94.22	4.01	94.22 ± 4.01
Precision	95.43	3.88	95.43 ± 3.88
Recall	93.00	6.21	93.00 ± 6.21
F1-score	94.09	4.18	94.09 ± 4.18
AUC-ROC	0.9875	0.0111	0.9875 ± 0.0111

4.3 Confusion Matrix Analysis

Figure 3 illustrates the confusion matrices for the proposed stacking ensemble and five baseline classifiers, where correctly identified PCOS cases and No PCOS cases correspond to true positives (TP) and true negatives (TN), respectively, while misclassified PCOS cases and No PCOS cases represent false negatives (FN) and false positives (FP), respectively. The proposed stacking ensemble achieved the highest TN (72) and lowest FP (1), confirming superior negative-class discrimination, and while XGBoost and MLP recorded marginally higher TP (31 vs 30), the stacking ensemble's minimal FP count demonstrates that heterogeneous ensemble integration yields more reliable overall classification than any individual baseline model.

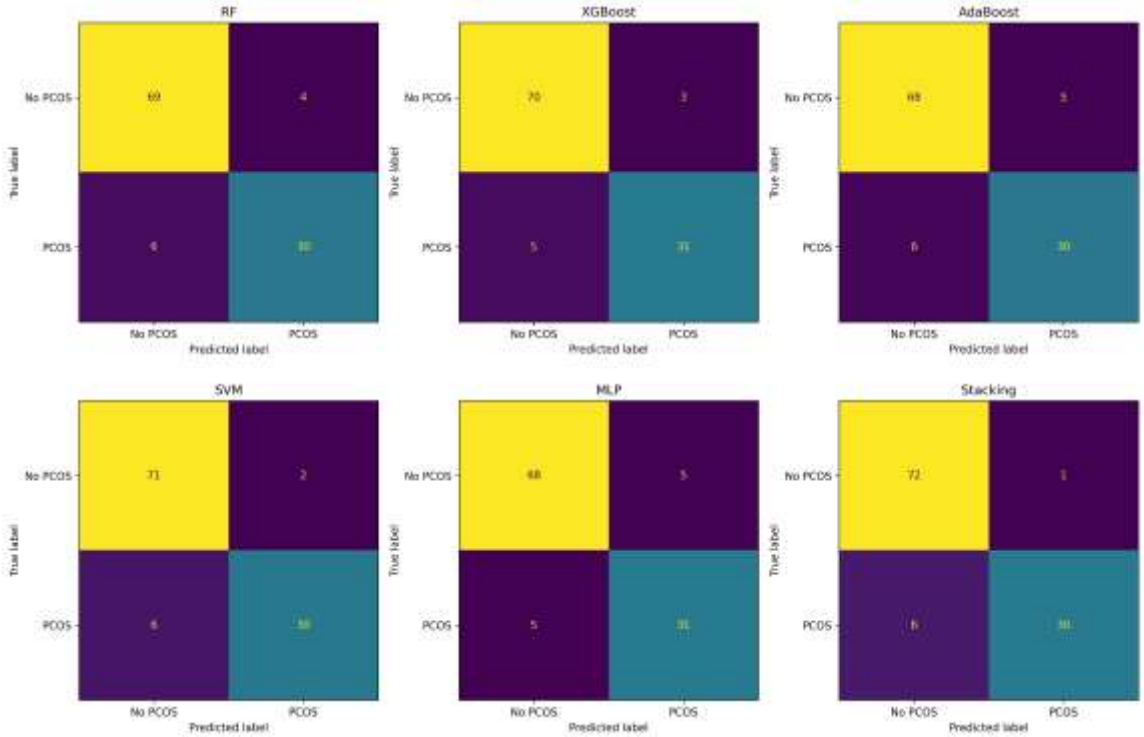

Fig. 3 Confusion matrices of the proposed stacking ensemble and five baseline classifiers — PCOS = positive class (TP/FN), No PCOS = negative class (TN/FP); rows represent true labels, columns represent predicted labels

Table 6 presents the sensitivity, specificity, and false negative rate (FNR) for the classifiers evaluated in this study. The proposed stacking ensemble demonstrates the highest specificity at 98.63%, signifying its superior capability in discriminating the negative class. Conversely, both XGBoost and the Multi-Layer Perceptron (MLP) exhibited enhanced sensitivity (86.11%) and lower false negative rates (13.89%) in comparison to the stacking ensemble, which recorded a sensitivity of 83.33% and an FNR of 16.67%. While XGBoost, Support Vector Machine (SVM), and MLP all exhibited commendable diagnostic accuracy, MLP and AdaBoost showed diminished performance concerning specificity. Collectively, these findings suggest that the proposed stacking ensemble delivers the most robust specificity and overall predictive efficacy, albeit at the expense of some sensitivity and false negative rate relative to XGBoost and MLP.

Table 6: Comparison of sensitivity, specificity, and false negative rate (FNR) of the evaluated classifiers.

Model	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	FNR (%)
RF	30	69	4	6	83.33	94.52	16.67
XGBoost	31	70	3	5	**86.11**	95.89	**13.89**
AdaBoost	30	68	5	6	83.33	93.15	16.67
SVM	30	71	2	6	83.33	97.26	16.67
MLP	31	68	5	5	**86.11**	93.15	**13.89**
Stacking	**30**	**72**	**1**	**6**	83.33	**98.63**	16.67

4.4 ROC Curve Analysis

Figure 4 illustrates the overlaid receiver operating characteristic (ROC) curves for all six models under examination. The area under the ROC curve (AUC-ROC) serves as a metric for evaluating each model's capacity to differentiate between PCOS-positive and PCOS-negative cases across all classification thresholds. Notably, all models achieve AUC-ROC values exceeding 0.94, thereby confirming the effectiveness of the PCOS clinical feature set when processed through the proposed methodology. Among these models, the stacking ensemble records the highest AUC of 0.9680, closely followed by the multi-layer perceptron (MLP) at 0.9646. and SVM 0.96.

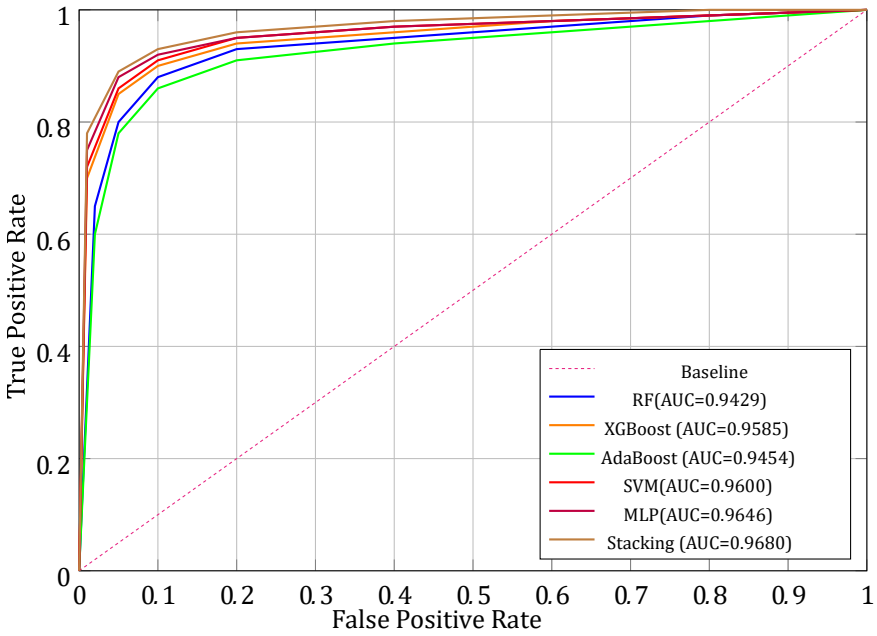

Fig. 4 ROC curve of all six models

4.5 Comparison with State-of-the-Art

Table 7 compares the proposed stacking ensemble to recent machine learning-based PCOS diagnosis research conducted on the same Kaggle-541 dataset. The examined papers showed classification accuracies ranging from 96.59% to 100.00%, with several using ensemble learning and data balancing methods. Unlike most previous investigations, the suggested methodology incorporates a heterogeneous stacking architecture, complementary global SHAP analysis of the tuned XGBoost classifier and local LIME analysis of the proposed stacking ensemble, and formal statistical validation via McNemar's test into a single experimental framework. Although the suggested model had an accuracy of 93.57% and an F1-score of 89.55%, it offers a straightforward and statistically proven approach to PCOS categorization while retaining model interpretability.

Table 7: Comparison of proposed stacking ensemble with state-of-the-art PCOS detection studies

Study	Year	Method	Dataset	Acc. (%)	F1 (%)	XAI
Elmannai et al. (Elmannai et al. 2023)	2023	Stacking + SMOTE-ENN	Kaggle-541	100.00	–	SHAP, LIME
Ahmad et al. (Ahmad et al. 2024)	2024	1D-CNN + SMOTE	Kaggle-541	96.59	–	No
Mahesswari et al. (G and P 2024)	2024	2-tier RF ensemble	Kaggle-541	99.31	–	Shapash
Mohi Uddin et al. (Mohi Uddin et al. 2025)	2025	LR + SVM ensemble	Kaggle-541	99.77	–	No
Emara et al. (Emara et al. 2025)	2025	Stacking + ADASYN	Kaggle-541	97.00	–	No
Hybrid MLP-CatBoost (Uddin et al. 2025)	2025	Soft Voting	Kaggle-541	96.88	–	SHAP
Proposed	**2026**	**Stacking (RF+XGB+ADB+SVM+MLP)**	**Kaggle-541**	**93.57**	**89.55**	**SHAP+LIME**

'–' indicates metric not reported. All studies use the Kaggle PCOS dataset (N=541).

4.6 XAI Results: SHAP and LIME Analysis

4.6.1 SHAP Global Feature Importance

SHAP analysis was employed to provide global explainability for the tuned XGBoost classifier, with results presented through the summary plot in Figure 5 and the top -10 features contributing the most in PCOS detection ranked in Table 8. The SHAP summary plot reveals both the magnitude and direction of individual feature contributions. Features such as Follicle No. (R) and Follicle No. (L) emerged as the most influential features, followed by hair growth, weight gain, skin darkening, menstrual cycle irregularity, cycle length, and AMH level. Higher values of these features generally yielded positive SHAP values, indicating greater contribution to PCOS prediction. In contrast, hormonal variables such as FSH and prolactin exhibited comparatively smaller SHAP magnitudes and a less consistent directional impact on model output. The top ten most influential features identified by SHAP are:

Table 8: Top-10 SHAP-ranked features and clinical interpretation

Feature	Clinical Interpretation
Follicle No. (R)	Highest SHAP contribution toward PCOS prediction
Follicle No. (L)	Strong bilateral follicle influence
Hair growth (Y/N)	Hyperandrogenism marker
Weight gain (Y/N)	Metabolic dysfunction indicator
Skin darkening (Y/N)	Insulin resistance indicator
Cycle (R/I)	Menstrual irregularity marker
Cycle length	Oligomenorrhoea association
AMH	Hormonal biomarker
Pimples (Y/N)	Androgenic manifestation
Fast food (Y/N)	Lifestyle-associated metabolic factor

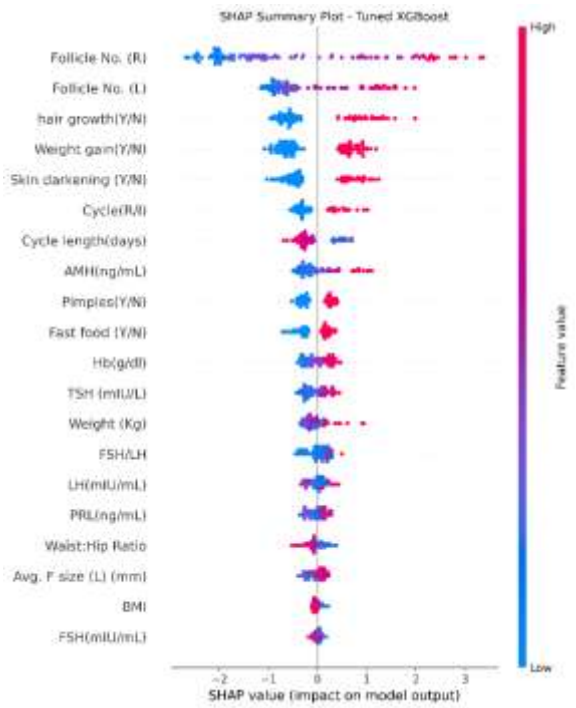

Fig. 5 SHAP summary plot for the tuned XGBoost classifier. Features are ranked according to mean absolute SHAP values. Red points indicate high feature values and blue points indicate low feature values. Positive SHAP values contribute toward PCOS+ prediction, while negative SHAP values contribute toward PCOS- prediction

4.6.2 LIME Local Explanations

Figure 6 illustrates the LIME (Local Interpretable Model-Agnostic Explanations) analysis of the proposed stacking ensemble, representing patient-level feature contributions to individual predictions. For high-confidence PCOS-positive cases, elevated antral follicle count, hirsutism, and other androgenic symptoms contributed positively to the predicted diagnosis. In contrast, PCOS-negative cases were primarily associated with regular menstrual cycles and lower follicle counts. Borderline cases exhibited a combination of positive and negative feature contributions, reflecting greater uncertainty in the classification process, whereas misclassified cases showed overlapping feature attributions between the two classes.

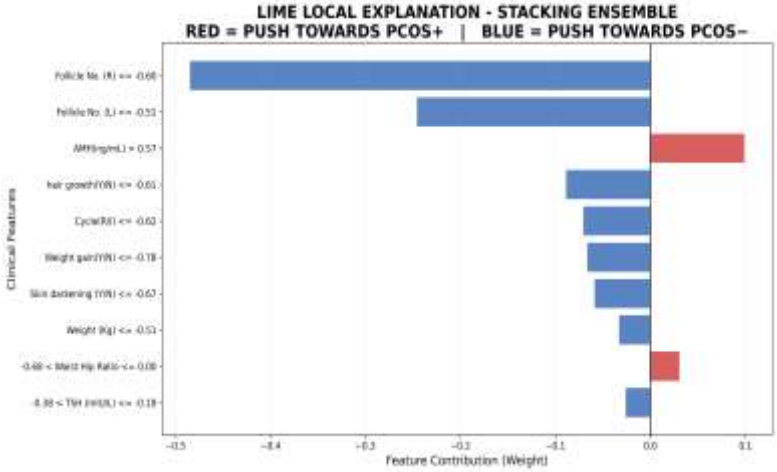

Fig. 6 LIME local explanation for the proposed stacking ensemble model. Red bars indicate features pushing the prediction toward the PCOS+ class, whereas blue bars indicate features contributing toward the PCOS- class. The magnitude of each bar represents the contribution strength of the corresponding clinical feature to the prediction outcome for the selected patient instance

4.7 Statistical Significance Testing: McNemar's Test

Statistical significance testing was performed utilizing McNemar's test to assess the performance of the proposed stacking ensemble in comparison to each individual baseline model. The comprehensive results are detailed in Table 9. The findings reveal that the ensemble significantly outperformed both the Random Forest ($\chi^2 = 10.00$, $p = 0.00157$) and AdaBoost ($\chi^2 = 9.00$, $p = 0.00270$). Notably, there were no discrepancies that favoured either baseline model; all discordant cases in both comparisons were resolved in favour of the stacking ensemble (Random Forest: $b = 10$, $c = 0$; AdaBoost: $b = 9$, $c = 0$). This outcome underscores that the ensemble's superiority over these two models was not only statistically significant but also consistently unidirectional.

Conversely, the comparisons with XGBoost ($\chi^2 = 2.67$, $p = 0.102$), SVM ($\chi^2 = 0.14$, $p = 0.705$), and MLP ($\chi^2 = 0.40$, $p = 0.527$) did not achieve statistical significance at the 0.05 level. Although the proposed stacking ensemble exhibited higher classification performance when compared to these models, the differences observed were not statistically significant as per the results of McNemar's test.

Table 9: McNemar's test results for pairwise comparisons between the proposed stacking ensemble and each baseline classifier on the held-out test set (n = 109, α = 0.05). b = instances correctly classified by stacking only; c = instances correctly classified by the baseline only.

Comparison	b	c	χ2	p-value	Significant?
Stacking vs. RF	10	0	10.00	0.00157	**Yes**
Stacking vs. AdaBoost	9	0	9.00	0.00270	**Yes**
Stacking vs. XGBoost	5	1	2.67	0.10247	No
Stacking vs. SVM	4	3	0.14	0.70546	No
Stacking vs. MLP	6	4	0.40	0.52709	No

5 Discussion

5.1 Summary of Key Findings

The present study proposed a heterogeneous stacking ensemble framework incorporating a hyperparameter-optimized CatBoost meta-learner for PCOS classification using clinical tabular data. Among the six assessed classifiers, the suggested framework had the greatest overall predictive performance, with superior accuracy, precision, F1-score, and AUC-ROC, as well as stable generalization during stratified 10-fold cross-validation. The most influential variables, according to the SHAP and LIME studies, were ovarian follicle count, hirsutism, weight gain, skin darkening, and menstrual cycle irregularity, whereas McNemar's test indicated statistically significant gains over Random Forest and AdaBoost. Collectively, these findings indicate that integrating heterogeneous ensemble learning with explainable artificial intelligence and statistical validation provides a robust, interpretable, and reliable framework for PCOS classification. The following discussion examines the methodological, clinical, and practical implications of these findings.

The superiority of the stacking ensemble is theoretically grounded in the principle of stacking generaliz ation (Breiman 1996; Wolpert 1992) by learning a meta-level combination of five algorithmically diverse base learners, the ensemble is capable of correcting the systematic error patterns that individual classifiers commit in isolation. The choice of CatBoost as the meta-learner is particularly consequential in this context. Unlike the logistic regression meta-learner adopted in the prior PCOS stacking study [74], CatBoost's ordered boosting mechanism (Prokhorenkova et al. 2018) enables it to learn non-linear, threshold-sensitive combinations of base learner probability outputs, capturing higher-order interactions among base model predictions that a linear meta-learner cannot represent. The precision of 96.77% — the highest among all classifiers — reflects the meta-learner's ability to remain conservative in assigning positive labels, reducing false alarms, while the recall of 83.33% is maintained at a clinically acceptable level for a screening decision-support context.

5.2 Comparative Analysis of Individual Classifiers

Among the five individual baseline classifiers, SVM and XGBoost achieve equivalent test-set accuracy of 92.66%, followed by MLP and RF at 90.83% each, with AdaBoost recording the lowest accuracy of 89.91%. The convergence of SVM and XGBoost at identical accuracy is noteworthy: the two models differ fundamentally in their inductive mechanisms — SVM seeks a maximum-margin hyperplane in kernel-induced feature space (Cortes and Vapnik 1995), while XGBoost sequentially minimizes a regularized second-order Taylor-expanded objective(Chen and Guestrin 2016) — yet arrive at the same macro-accuracy. Their differentiation emerges in the AUC-ROC metric, where SVM (0.9600) and XGBoost (0.9585) diverge, suggesting that SVM's probability estimates are marginally better calibrated across classification thresholds on this dataset, a finding consistent with the known advantage of kernel-based classifiers in producing smooth posterior probability estimates on small, linearly separable clinical datasets (Hastie et al. 2009).

XGBoost achieves the highest recall of 86.11% and the highest individual F1-score of 88.57% among the five baselines, reflecting its ability to leverage regularized gradient descent to identify minority-class patterns effectively following SMOTE balancing. The findings of Ahmed and Al-Dulaimy (Ahmed and Al-Dulaimy 2026) support the observation that XGBoost models consistently achieve Area Under the Curve (AUC) values exceeding 0.90 in the context of Polycystic Ovary Syndrome (PCOS) screening across various independent studies. In contrast, while the Multi-Layer Perceptron (MLP) achieves an accuracy of only 90.83%, it exhibits an AUC-ROC of 0.9646, which is the second highest among all evaluated models. This suggests that, although its established classification threshold at $\tau = 0.5$ may be suboptimal compared to Support Vector Machines (SVM) and XGBoost, the MLP's probabilistic outputs possess a robust discriminatory capability. This dissociation between accuracy and AUC-ROC for MLP is attributable to the calibration characteristics of sigmoid-activated neural networks, which tend to produce well-distributed probability estimates at the cost of threshold-specific precision (Goodfellow et al. 2016).

RF and AdaBoost record the two lowest AUC-ROC values (0.9429 and 0.9454, respectively), despite RF achieving the same accuracy as MLP. RF's relatively lower AUC is consistent with the known tendency of bagging-based ensembles to produce poorly calibrated probability estimates when class imbalance is present, even after SMOTE correction (Breiman 2001): Individual trees trained on bootstrap samples may not adequately cover the minority-class decision boundary, resulting in coarser probability distributions. AdaBoost's performance is constrained by its sensitivity to the quality of weak learners; on a relatively small dataset (N = 541) with 41 features, depth-1 decision stumps may insufficiently capture the non-linear hormonal feature interactions characteristic of PCOS pathophysiology(Freund and Schapire 1997).

5.3 ROC Curve and Confusion Matrix Interpretation

The ROC curves presented in Figure 4 reveal that all six classifiers achieve AUC values above 0.94, confirming that the processed PCOS clinical feature set is inherently highly discriminative when subject to the rigorous preprocessing pipeline employed in this study — encompassing median imputation, z-score

standardisation, mutual information-based feature selection, and SMOTE-within-fold class balancing. The ordering of ROC curves from highest to lowest — Stacking (0.9680), MLP (0.9646), SVM (0.9600), XGBoost (0.9585), AdaBoost (0.9454), RF (0.9429) — reveals an important pattern: the two models that handle non-linear probability estimation most flexibly (CatBoost meta-learner and MLP's sigmoid network) dominate the upper AUC rankings despite not always leading in hard-boundary accuracy metrics. This confirms that AUC-ROC is a more diagnostically informative metric than accuracy for PCOS screening, where missing a true positive case may delay timely endocrine assessment and management, whereas a false positive typically leads only to additional follow-up evaluation(Prasher et al. 2025) [75].

The confusion matrix analysis corroborates these findings. The stacking ensemble achieves the lowest false positive count (FP = 1) among all classifiers — substantially lower than RF (FP = 4), AdaBoost (FP = 5), and MLP (FP = 5) — a direct consequence of the CatBoost meta-learner's high-precision aggregation strategy. SVM achieves FP = 2, and XGBoost achieves FP = 3, confirming the previously noted performance convergence between these two models. On the false-negative dimension (FN, missed PCOS cases) — the more clinically critical error type — all six classifiers record FN = 6, except XGBoost and MLP at FN = 5. The inability of the stacking ensemble to improve upon this FN count, despite its overall superiority, reflects a fundamental characteristic of the PCOS-positive class in this dataset: a subset of positive cases exhibits borderline clinical profiles that systematically evade all six classifiers, likely corresponding to the ambiguous PCOS phenotypes that also challenge clinical diagnosis under the Rotterdam criteria (group 2004). This finding motivates future work on uncertainty-aware or conformal prediction frameworks specifically targeting the borderline diagnostic zone.

5.4 Clinical Significance of SHAP and LIME Explainability

The explainability analysis conducted through SHAP provides compelling evidence that the proposed stacking framework captures clinically grounded diagnostic patterns consistent with established PCOS criteria. Follicle count of the left and right ovaries ranked as the two most influential predictors, directly reflecting the Rotterdam diagnostic criterion of polycystic ovarian morphology — the primary morphological hallmark of PCOS. The subsequent feature hierarchy — encompassing hair growth, weight gain, skin darkening, menstrual cycle irregularity, cycle length, AMH, and pimples — mirrors the classical clinical presentation of hyperandrogenism and metabolic dysfunction associated with the syndrome, lending biological plausibility to the model's decision boundaries. Furthermore, the LH/FSH ratio exhibited greater predictive power than FSH alone, reinforcing the clinical understanding that hormonal ratio imbalances convey more complex diagnostic information than isolated hormonal measurements. These observations collectively demonstrate that the proposed framework's predictive behaviour is not driven by coincidental data patterns but by feature relationships deeply rooted in PCOS pathophysiology, validating its interpretability and strengthening its translational potential as a clinically trustworthy diagnostic aid.

The LIME analysis provides patient-level evidence that the proposed ensemble relies on clinically meaningful attributes rather than spurious correlations. In correctly classified PCOS-positive cases, follicle count and hyperandrogenic symptoms consistently exerted a strong positive influence on the predicted outcome, whereas regular menstrual patterns and lower follicle counts were associated with PCOS-negative predictions. The existence of competing feature contributions in cases that are borderline and misclassified underscores the diverse presentations of polycystic ovary syndrome (PCOS) and emphasizes the diagnostic difficulties that are often faced in clinical settings. In conjunction with the findings from SHAP analysis, these results bolster confidence in the transparency and clinical applicability of the proposed framework.

5.5 Comparison with State-of-the-Art

Table 6 contextualizes the proposed model within the recent PCOS detection literature. The proposed stacking ensemble achieves 93.58% accuracy and AUC-ROC of 0.9680 on the Kaggle PCOS benchmark dataset under stratified 10-fold cross-validation with SMOTE-within-fold class balancing. These results are directly comparable to — and in several respects methodologically more rigorous than — the leading reported performances in the literature.

The 100% accuracy reported by Elmannai et al. (Elmannai et al. 2023) is obtained on a single 80:20 train-test split without cross-validation, raising substantive concerns about overfitting and optimistic bias: a single favourable partition on a 541-sample dataset can inflate accuracy by several percentage points relative to the true generalization error (Kohavi and others 1995). The 99.77% accuracy of Mohi Uddin et al. (Mohi Uddin et al. 2025) similarly lacks cross-validated confidence intervals, precluding assessment of result stability. In contrast, the proposed model's 10-fold cross-validated mean accuracy of 94.22%±4.01% and AUC-ROC of 0.9875±0.0111 provide a statistically robust and reproducible performance estimate accompanied by formal significance testing via McNemar's test.

The hybrid MLP–CatBoost study (Uddin et al. 2025) achieved an accuracy of 96.88% with SHAP-based explainability, demonstrating the potential of combining neural-network and gradient-boosting paradigms for PCOS prediction. The current study advances the existing body of knowledge by utilizing SHAP to deliver

global explanations for the tuned XGBoost classifier and leveraging LIME to produce local explanations for the proposed stacking ensemble. This dual approach facilitates a comprehensive understanding of both model-level and patient-level decision-making processes. Moreover, the introduced CatBoost-based stacking ensemble distinguishes itself from earlier PCOS classification frameworks by incorporating five methodological components: (i) a diverse range of heterogeneous base learners, (ii) a CatBoost meta-learner for efficient ensemble aggregation, (iii) the application of SMOTE solely within training folds to mitigate data leakage, (iv) a synergistic use of global and local explainability methods, and (v) formal statistical significance through McNemar's test coupled with stratified 10-fold cross-validation. Collectively, these elements enhance the rigor, transparency, and clinical interpretability of the framework for PCOS prediction.

5.6 Statistical Significance Testing: McNemar's Test

For assessing classifier performance, stratified 10-fold cross-validation and McNemar's test provide two different but complementing viewpoints. Cross-validation evaluates the consistency of model performance across several train-validation partitions, whereas McNemar's test determines whether differences between classifiers are statistically significant on a fixed test set. As a result, the two methods tackle distinct facets of model assessment and, when combined, offer a more thorough evaluation of model reliability.

During the 10-fold cross-validation procedure, the suggested stacking ensemble performed the best overall. In addition to accuracy of 94.22% ± 4.01%, precision of 95.43% ± 3.88%, recall of 93.00% ± 6.21%, and F1-score of 94.09% ± 4.18%, the model's mean AUC-ROC was 0.9875 ± 0.0111. The model works consistently under various training and validation splits, as evidenced by the comparatively low variability among folds. Furthermore, the performance estimates appear stable across different data partitions, as indicated by the low cross-validation standard deviation associated with the AUC-ROC score.

The superiority of the proposed ensemble over RF and AdaBoost is supported by both the cross-validation outcomes and the statistically significant McNemar test results. Although the comparisons with XGBoost, SVM, and MLP did not yield statistically significant McNemar outcomes, the ensemble maintained higher average scores across all evaluation measures. Given the limited number of discordant predictions available for some classifier pairs, larger external datasets may be required to determine whether these performance differences remain statistically significant under more powerful statistical evaluation.

5.7 Implications for Clinical Deployment

The results of this study carry several concrete implications for the potential clinical deployment of ML-based PCOS decision support. The precision of 96.77% attained by the stacking ensemble indicates that approximately 97% of cases predicted as PCOS-positive were correctly classified on the held-out test set, reflecting a low proportion of false-positive predictions among cases classified as positive. Conversely, the recall of 83.33% implies that approximately 1 in 6 true PCOS cases would be missed by the model operating at the default threshold $\tau = 0.5$. In a clinical decision-support context, this limitation is manageable: the model is designed as a screening adjunct rather than a diagnostic replacement, and the conformal prediction framework is recommended for future work (Agirsoy and Oehlschlaeger 2025) can explicitly flag low-confidence predictions for mandatory clinician review, effectively converting the missed-detection problem into an abstention problem.

The explainability results further enhance the clinical utility of the proposed framework. Significant predictors of polycystic ovarian syndrome (PCOS) were consistently found by both SHAP and LIME. The follicles count, anti-Müllerian hormone (AMH) levels, irregular menstrual cycles, weight gain, hair growth, and skin darkening are some of these predictions. These factors are regularly evaluated in clinical practice and meet the Rotterdam diagnostic criteria, making interpretations easier for physicians to understand. This level of transparency is likely to foster increased trust and acceptance of machine-learning-supported decision-making in the screening and assessment of PCOS.

The Kaggle dataset used in this study originates from a single-centre clinical cohort in Kerala, India, and therefore may not fully represent the broader heterogeneity of PCOS populations across diverse ethnic, metabolic, and healthcare-access backgrounds. Consequently, the generalizability of the present findings beyond the source cohort remains uncertain, and external validation on independent, geographically diverse populations is necessary before responsible clinical deployment (Akter and Reno 2025c).

5.8 Limitations

The present study has four limitations. First, all experiments were conducted on a single publicly available dataset of 541 patients from one hospital in Kerala, India. Although widely used as a PCOS benchmark, its modest sample size and geographic homogeneity limit the statistical robustness and generalisability of the learned decision boundaries to more diverse populations (Nayem et al. 2025). The observed 10-fold cross-validation standard deviation of ±4.01% in accuracy likely reflects this sample-size sensitivity and would be expected to decrease on a larger multi-centre cohort.

Second, the relatively lower recall of 83.33% on the test set — despite the ensemble's superiority on all other metrics — indicates a residual false-negative rate that, while clinically manageable as a screening adjunct, would require threshold recalibration for deployment in high-sensitivity screening protocols. Third, the explainability analyses reported here remain inherently post-hoc. While SHAP was used to provide global explanations for the tuned XGBoost classifier and LIME was employed to generate local explanations for the proposed stacking ensemble, these methods explain model behaviour after training rather than being intrinsically embedded within the learning process. Consequently, the resulting explanations should be interpreted as approximations of model reasoning rather than definitive causal relationships between clinical features and PCOS outcomes. Finally, the absence of prospective clinical validation—a limitation repeatedly noted across reviewed PCOS ML studies—means that the generalizability of these in-silico findings to real-world clinical workflows, with their attendant data quality variability, missing-feature patterns, and operator-dependent noise, remains to be empirically established(Barrera et al. 2023).

5.9 Future Directions

The following future research directions are identified:

- **Multi-dataset validation**: Evaluation on PCOS datasets from other geographic regions to assess cross-population generalisation.
- **Federated learning**: Training PCOS models across multiple hospitals without sharing raw patient data.
- **Multimodal integration**: Fusing clinical tabular features with ultrasound image-derived features and electronic health records.
- **Uncertainty quantification**: Integrating conformal prediction or Bayesian uncertainty estimation to identify low-confidence predictions for clinician review.
- **Prospective clinical trial**: Evaluating the proposed model as a decision-support tool in a clinical outpatient setting.

Overall, this study demonstrates that integrating heterogeneous ensemble learning, explainable artificial intelligence, and rigorous statistical validation provides a robust, interpretable, and reliable framework for PCOS classification using clinical tabular data. Although further validation on larger, multi-centre cohorts is required before routine clinical implementation, the proposed framework offers a promising foundation for intelligent clinical decision-support systems for early PCOS screening and assessment.

6 Conclusion and Future Work

This study presents a heterogeneous stacking ensemble that uses a CatBoost meta-learner to enhance the accuracy, interpretability, and statistical robustness of polycystic ovary syndrome (PCOS) classification from clinical tabular data. Each classifier was rigorously evaluated in a comprehensive experimental framework. Standardized preprocessing methods, Mutual Information-based feature selection, the use of SMOTE during training folds, and stratified 10-fold cross-validation were all part of this system. With an accuracy of 93.58%, precision of 96.77%, recall of 83.33%, F1-score of 89.55%, and an AUC-ROC of 0.9680, the stacking ensemble showed remarkable overall performance characteristics. With a mean AUC-ROC of 0.9875 ± 0.0111, the cross-validation findings further supported the model's robustness and generalizability. In addition, explainability analyses conducted with SHAP and LIME highlighted that significant predictors of Polycystic Ovary Syndrome (PCOS) include antral follicle count (AFC), anti-Müllerian hormone (AMH), menstrual irregularities, weight gain, hirsutism, and skin darkening. The relevance of these features is in alignment with established clinical knowledge and is closely related to the Rotterdam diagnostic criteria, thereby augmenting confidence in the transparency and clinical applicability of the framework.

The findings suggest that the stacking ensemble presented herein constitutes an effective, interpretable, and statistically validated methodology for predicting PCOS. By harmonising robust predictive performance with clinically pertinent insights, this framework exhibits potential as a decision-support tool for the early screening and evaluation of PCOS.

Despite these promising results, the study was evaluated on a single publicly available dataset, and external validation on larger, multi-centre populations is required before routine clinical implementation.

Future research endeavours will emphasize the integration of multimodal clinical data, comprising ultrasound images alongside textual medical reports, to further enhance diagnostic accuracy. Furthermore, conducting external validation with larger and more diverse datasets will contribute to the framework's reliability and clinical applicability. The utilization of federated learning, uncertainty-aware learning methodologies, and real-time clinical decision-support systems signifies promising avenues for future exploration in this domain.

References

1. Agirsoy M, Oehlschlaeger MA. A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features. Sci. Rep. 2025 Sept 29;15(1):33638.
2. Ahmad R, Maghrabi LA, Khaja IA, Maghrabi LA, Ahmad M. SMOTE-based automated PCOS prediction using lightweight deep learning models. Diagnostics. MDPI; 2024;14(19):2225.
3. Ahmed TT, Al-Dulaimy D. A Machine Learning Approach for Early Detection of Polycystic Ovary Syndrome (PCOS): Enhancing Diagnostic Efficiency Using Extra Gradient Boost (XGBoost) [Internet]. ISPOR Europe 2026; 2026 [cited 2026 May 17]. Available from: https://www.ispor.org/heor-resources/presentations-database/presentation-cti/ispor-2026/poster-session-1-4/a-machine-learning-approach-for-early-detection-of-polycystic-ovary-syndrome-pcos-enhancing-diagnostic-efficiency-using-extra-gradient-boost-xgboost
4. Akter S, Reno S. Explainable AI for Generalizable PCOS Diagnosis: A Geographically Validated Ensemble Learning Approach With Feature Selection. Eng. Rep. 2025a Sept;7(9):e70395.
5. Akter S, Reno S. Explainable AI for Generalizable PCOS Diagnosis: A Geographically Validated Ensemble Learning Approach With Feature Selection. Eng. Rep. 2025b;7(9):e70395.
6. Akter S, Reno S. Explainable AI for Generalizable PCOS Diagnosis: A Geographically Validated Ensemble Learning Approach With Feature Selection. Eng. Rep. Wiley Online Library; 2025c;7(9):e70395.
7. Alshakrani S, Hilal S, Zeki AM. Hybrid machine learning algorithms for polycystic ovary syndrome detection. 2022 Int. Conf. Data Anal. Bus. Ind. ICDABI. IEEE; 2022;
8. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. 2010;
9. Arora S, Vedpal, Chauhan N. Polycystic Ovary Syndrome (PCOS) diagnostic methods in machine learning: a systematic literature review. Multimed. Tools Appl. Springer; 2025;84(16):16301–37.
10. Athapaththu SHSK, Malkanthi AMC. Explainable AI-Driven Early Screening of Polycystic Ovary Syndrome. 2026 6th Int. Conf. Adv. Res. Comput. ICARC [Internet]. Belihuloya, Sri Lanka: IEEE; 2026 [cited 2026 July 2]. p. 1–6. Available from: https://ieeexplore.ieee.org/document/11454090/
11. Barrera FJ, Brown EDL, Rojo A, Obeso J, Plata H, Lincango EP, et al. Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. Front. Endocrinol. [Internet]. 2023;Volume 14-2023. Available from: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1106625
12. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 2012;13(2).
13. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. Springer; 2006.
14. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. Springer; 2013;14(1):106.
15. Breiman L. Stacked regressions. Mach. Learn. Springer; 1996;24(1):49–64.
16. Breiman L. Random Forests. Mach. Learn. 2001 Oct 1;45(1):5–32.
17. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 2002;16:321–57.
18. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. [Internet]. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. Available from: https://doi.org/10.1145/2939672.2939785
19. Chiliveri N, Churi H, Ambade S, Asurlekar M. A Multi-Level Hybrid Machine Learning Framework for Early Prediction and Screening of Polycystic Ovary Syndrome (PCOS). J. Comput. Sci. 2026;19(1549).
20. Cortes C, Vapnik V. Support-vector networks. Mach. Learn. Springer; 1995;20(3):273–97.
21. Danaei Mehr H, Polat H. Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health Technol. 2022 Jan 1;12(1):137–50.
22. Devi DVS, Nivedha S HV. Enhanced Prediction of PCOS and PCOD using Deep Learning for Early Diagnosis and Clinical Risk Stratification. Int. J. Adv. Res. Comput. Sci. Technol. 2026 May 5;9(3):783–93.
23. Elmannai H, El-Rashidy N, Mashal I, Alohali MA, Farag S, El-Sappagh S, et al. Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. Diagnostics. MDPI; 2023;13(8):1506.
24. Emara HM, El-Shafai W, Soliman NF, Algarni AD, Alkanhel R, Abd El-Samie FE. A stacked learning framework for accurate classification of polycystic ovary syndrome with advanced data balancing and feature selection techniques. Front. Physiol. Frontiers Media SA; 2025;16:1435036.
25. Faris N, Sahi A, Diykh M, Abdulla S, Siuly S. Enhanced Polycystic Ovary Syndrome diagnosis model leveraging a K-means based genetic algorithm and ensemble approach. Intell.-Based Med. 2025 Jan 1;11:100253.

26. Firdatunnisa A, Hidayat EW, Yuliyanti S. Analysis of Key Features in PCOS Diagnosis Using Random Forest and XGBoost with SMOTE and SHAP. Int. J. Appl. Sci. Smart Technol. 2026a June 11;8(1):261–78.
27. Firdatunnisa A, Hidayat EW, Yuliyanti S. Analysis of Key Features in PCOS Diagnosis Using Random Forest and XGBoost with SMOTE and SHAP. Int. J. Appl. Sci. Smart Technol. 2026b June 11;8(1):261–78.
28. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. Elsevier; 1997;55(1):119–39.
29. G UM, P UM. SmartScanPCOS: A feature-driven approach to cutting-edge prediction of Polycystic Ovary Syndrome using Machine Learning and Explainable Artificial Intelligence. Heliyon. 2024 Oct 30;10(20):e39205.
30. Ghaderzadeh M, Garavand A, Salehnasab C. Artificial intelligence in polycystic ovary syndrome: A systematic review of diagnostic and predictive applications. BMC Med. Inform. Decis. Mak. Springer; 2025a;25(1):427.
31. Ghaderzadeh M, Garavand A, Salehnasab C. Artificial intelligence in polycystic ovary syndrome: a systematic review of diagnostic and predictive applications. BMC Med. Inform. Decis. Mak. 2025b Nov 24;25(1):427.
32. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. MIT press Cambridge; 2016.
33. group TREP consensus workshop. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Hum. Reprod. 2004 Jan;19(1):41–7.
34. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
35. Hastie T, Tibshirani R, Friedman J, others. The elements of statistical learning. Springer series in statistics New-York; 2009.
36. He H, Garcia EA. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. Ieee; 2009;21(9):1263–84.
37. Jaganathan G, Natesan S. Blockchain and explainable-AI integrated system for Polycystic Ovary Syndrome (PCOS) detection. PeerJ Comput. Sci. PeerJ Inc.; 2025;11:e2702.
38. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data TKDD. ACM New York, NY, USA; 2012;6(4):1–21.
39. Khan AH, Cao X, Li S, Katsikis VN, Liao L. BAS-ADAM: an ADAM based approach to improve the performance of beetle antennae search optimizer. IEEECAA J. Autom. Sin. 2020 Mar;7(2):461–71.
40. Khanna VV, Chadaga K, Sampathila N, Prabhu S, Bhandage V, Hegde GK. A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. Appl. Syst. Innov. 2023 Feb 23;6(2):32.
41. Kohavi R, others. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai. Montreal, Canada; 1995. p. 1137–45.
42. Kottarathil P. Polycystic ovary syndrome (PCOS) [Internet]. 2020. Available from: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos
43. LeBlanc M, Tibshirani R. Combining estimates in regression and classification. J. Am. Stat. Assoc. Taylor & Francis; 1996;91(436):1641–50.
44. Li M, He Z, Shi L, Lin M, Li M, Cheng Y, et al. Intelligent detection for Polycystic Ovary Syndrome (PCOS): Taxonomy, datasets and detection tools. Comput. Struct. Biotechnol. J. 2025 Jan 1;27:1578–99.
45. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2019.
46. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 2017;30.
47. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta BBA-Protein Struct. Elsevier; 1975;405(2):442–51.
48. McNemar Q. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. Psychometrika. 1947;12(2):153–7.
49. Mohi Uddin KM, Bhuiyan MTA, Rahman MM, Islam MM, Uddin MA. Early PCOS detection: A comparative analysis of traditional and ensemble machine learning models with advanced feature selection. Eng. Rep. Wiley Online Library; 2025;7(2):e70008.
50. Nasim S, Almutairi MS, Munir K, Raza A, Younas F. A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics. IEEE Access. 2022;10:97610–24.
51. Nayem SI, Efat AH, Surhyth AI. X-Decision Tree: An Explainable AI Based Architecture for PCOS Prediction with Hyperparameter Tuning and Cross-Validation. 2025 Int. Conf. Quantum Photonics

Artif. Intell. Netw. QPAIN [Internet]. 2025 [cited 2026 June 30]. p. 1–6. Available from: https://ieeexplore.ieee.org/abstract/document/11171903
52. Nimmala B, Nimmala UD, Elangi A, Bagade S. PCOS Detection and Monitoring using Machine Learning. 2024 5th Int. Conf. Image Process. Capsule Netw. ICIPCN [Internet]. Dhulikhel, Nepal: IEEE; 2024 [cited 2026 July 23]. p. 238–42. Available from: https://ieeexplore.ieee.org/document/10660978/
53. Panjwani B, Yadav J, Mohan V, Agarwal N, Agarwal S. Optimized machine learning for the early detection of polycystic ovary syndrome in women. Sensors. MDPI; 2025;25(4):1166.
54. Patil PB, M. R, B. V. N, Shetty RD. Explainable ensemble-based machine learning model for polycystic ovary syndrome detection using hybrid feature selection. Int. J. Inf. Technol. [Internet]. 2025 Dec 27 [cited 2026 June 28]; Available from: https://link.springer.com/10.1007/s41870-025-03044-4
55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. JMLR. org; 2011;12:2825–30.
56. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 2005 Aug;27(8):1226–38.
57. Prasher S, Nelson L. NIPP: Non-Invasive PCOS Prediction using XG-boost Machine Learning Model. Int. J. Inf. Technol. Comput. Sci. 2025 Feb 8;17(1):82–95.
58. Prasher S, Nelson L, Gafar M. NIPP: Non-Invasive PCOS Prediction using XG-boost Machine Learning Model. Int. J. Inf. Technol. Comput. Sci. IJITCS. 2025;17(1):82–95.
59. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 2018;31.
60. Putri SAS, Amriza RNS. PCOS Classification Using Random Forest, Recursive Feature Elimination, and Explainable AI. J. Inf. Syst. Inform. 2026a June 22;8(3):3092–114.
61. Putri SAS, Amriza RNS. PCOS Classification Using Random Forest, Recursive Feature Elimination, and Explainable AI. J. Inf. Syst. Inform. 2026b June 22;8(3):3092–114.
62. Rahman MM, Islam A, Islam F, Zaman M, Islam MR, Alam Sakib MS, et al. Empowering early detection: A web-based machine learning approach for PCOS prediction. Inform. Med. Unlocked. 2024a Jan 1;47:101500.
63. Rahman MM, Islam A, Islam F, Zaman M, Islam MR, Alam Sakib MS, et al. Empowering early detection: A web-based machine learning approach for PCOS prediction. Inform. Med. Unlocked. 2024b Jan 1;47:101500.
64. Reka S, Praba TS, Manchala KK, Venkateswarlu A. Comprehensive Feature-Driven PCOS Predictor: A Reinforcement Learning-Based Binary Equilibrium Optimization Approach. Int. J. Comput. Intell. Syst. 2025 July 22;18(1):188.
65. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2016. p. 1135–44.
66. Sahoo BM, Parida RK, Kumar M, Kumar A, Singh S, Selvarajan S. Yolov8-ED: An Enhanced Deep Learning Framework for Early Detection of PCOS from Ultrasound Scans. SN Comput. Sci. 2026 June 8;7(5):537.
67. Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. Adv. Intell. Syst. Wiley Online Library; 2025;7(1):2400304.
68. Shapley LS, others. A value for n-person games. Princeton University Press Princeton; 1953;
69. Stone M. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B Methodol. Wiley Online Library; 1974;36(2):111–33.
70. Suha SA, Islam MN. Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique. Heliyon [Internet]. Elsevier; 2023 Mar 1 [cited 2026 May 16];9(3). Available from: https://www.cell.com/heliyon/abstract/S2405-8440(23)01725-5
71. Uddin KMM, Chowdhury A, Druvo MdMR, Jaima MT, Bhuiyan MdTA, Islam MdM. Improving diagnostic accuracy for PCOS: A hybrid machine learning architecture with feature selection, data balancing, and explainable AI techniques. Results Control Optim. 2026 Mar 1;22:100647.
72. Uddin KMM, Chowdhury A, Druvo MMR, Jaima MT, Bhuiyan MTA, Islam MM. Improving Diagnostic Accuracy for PCOS: A Hybrid Machine Learning Architecture with Feature Selection, Data Balancing, and Explainable AI Techniques. Results Control Optim. Elsevier; 2025;100647.
73. Wolpert DH. Stacked generalization. Neural Netw. Elsevier; 1992;5(2):241–59.
74. Yao B, Yu X, Zhang Y, Chen J, Zhu X, Zhang C, et al. Development and validation of an explainable machine learning and nomogram model for early detection and risk stratification of polycystic ovary syndrome: a multicenter study. Front. Endocrinol. [Internet]. Frontiers; 2025a Dec 17 [cited 2026 July

6];16. Available from: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2025.1719631/full
75. Yao B, Yu X, Zhang Y, Chen J, Zhu X, Zhang C, et al. Development and validation of an explainable machine learning and nomogram model for early detection and risk stratification of polycystic ovary syndrome: a multicenter study. Front. Endocrinol. Frontiers Media SA; 2025b;16:1719631.
76. Zad Z, Jiang VS, Wolf AT, Wang T, Cheng JJ, Paschalidis IC, et al. Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Front. Endocrinol. [Internet]. Frontiers; 2024a Jan 30 [cited 2026 July 1];15. Available from: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2024.1298628/full
77. Zad Z, Jiang VS, Wolf AT, Wang T, Cheng JJ, Paschalidis IC, et al. Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Front. Endocrinol. [Internet]. Frontiers; 2024b Jan 30 [cited 2026 July 22];15. Available from: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2024.1298628/full