ENHANCED ML FRAMEWORK FOR PCOS DIAGNOSIS: OPTIMISED XGBOOST MODEL

Theepika[1], Santhana Megala[2], Mythili[3], Usha Devi[4], Balaganesh Duraisamy[5]

[1]Research Scholar, Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore & Assistant Professor, Department of Computer Applications, Dr. Mahalingam College of Engineering and Technology, Pollachi.
[2]Assistant Professor, Department of Computer Technology, Hindusthan College of Arts & Science, Coimbatore.
[3]Professor & Head, Department of Computer Technology, Hindusthan College of Arts & Science, Coimbatore.
[4]Assistant Professor, Department of Computer Science, Rathinam College of Arts and Science, Coimbatore, India.
[5]Lecturer of Faculty Computer Science and Multimedia, Berlin School of Business & Innovation (BSBI), Berlin, Germany.

ABSTRACT

Polycystic ovary syndrome (PCOS) is a very common endocrine disorder that affects women of child bearing age across the globe with a global prevalence of between 8-13 percent. Although it is a common disease, PCOS is highly undiagnosed, as the clinical picture is heterogeneous, and the diagnosis is based on costly hormonal tests and ultrasound studies. Later diagnosis causes chronic complications such as infertility, diabetes mellitus type 2, heart diseases, and endometrial cancer. This article reports a machine learning model of early and accurate prediction of PCOS based on an ensemble of Random Forest and XGBoost classifiers that will be trained on a multi-feature clinical dataset of hormonal measurements, anthropometric measurements, ultrasonographic results, and lifestyle factors. The dataset, which is obtained on the publicly accessible Kaggle PCOS dataset (n=539 patients, n=41 features), was preprocessed with a severe approach such as missing value imputation, label encoding, feature selection through Recursive Feature Elimination (RFE), and Synthetic Minority Oversampling Technique (SMOTE) to handle the imbalance in the classes. Experimental performance shows that XGBoost has the best performance with an accuracy of 92.8, F1-score of 0.922, and AUC-ROC of 0.971 whilst the random forest has an accuracy of 91.3 and AUC-ROC of 0.963 and is way better than the baseline classifiers such as logistic regression (81.2) and The most discriminative predictors based on the analysis of feature importance are follicle count, anti-Mullerian hormone (AMH) level, FSH:LH ratio, and cycle irregularity. The suggested system presents an efficient, non-invasive, and scaled clinical decision support tool in gynaecology practice.

KEYWORDS

Random Forest, XGBoost, Ensemble Learning, Clinical Decision Support, Feature Importance, Medical Machine Learning, PCOS Detection,

1. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a multifaceted hormonal disorder, which is defined by hyperandrogenism, ovulatory dysfunction, and polycystic morphology of the ovary. It is a condition that impacts one hundred and six million women across the world and is the major

DOI: 10.5121/ijait.2025.16301

cause of anovulatory infertility among women of reproductive age [3]. The most popular diagnostic system is the Rotterdam consensus in [13], where at least two out of three characteristics must be present: oligo/anovulation, clinical or biochemical evidence of hyperandrogenism, and polycystic ovaries on ultrasound. Yet, this heterogeneity implies that there is no biomarker that is adequate to make the diagnosis, and the condition has different manifestations in ethnic groups, body mass index (BMI) categories, and life stages.

Existing clinical diagnoses are based on serum hormone tests such as luteinising hormone (LH), follicle-stimulating hormone (FSH), testosterone, anti-Mullerian hormone (AMH) and fasting insulin in combination with pelvic ultrasound to quantify antral follicles. The multi-step diagnostic route is costly, time-consuming, and involves access by a specialist, which adds to a two-year or longer diagnostic delay among many patients [9]. Misdiagnosis and underdiagnosis are especially acute in low- and middle-income countries where specialists gynaecological care is scarce.

Machine learning (ML) offers even more effective alternative: having been trained on non-linear and complex patterns on multi-modal patient data, ML models can assist clinicians to identify the risk of PCOS based on regular clinical parameters, which may elevate at-risk patients to primary care status before they can be sent to specialists. Ensemble approaches with both Random Forest and XGBoost are especially suitable to this task because they are resistant to noisy and correlated features, feature importance scores are interpretable, and empirical results on tabular clinical data are very promising.

The primary findings of the paper are as follows:

- A clinical PCOS data end-to-end process that involved missing value imputation, categorical encoding, feature selection using RFE, and SMOTE oversampling.
- Stratified 80/20 train-test split Comparative evaluation of six ML classifiers, such as Logistic Regression,Naive Bayes, SVM, Decision Tree, Random Forest and XGBoost.
- Feature importance analysis of both random forest (Gini impurity) and XGBoost (gain-based) to determine the most clinically significant predictors of PCOS.
- Calibration of models in detail, ROC curve and confusion-matrix analysis to ascertain clinical utility, and sensitivity and specificity trade-offs, on the basis of a screening setting.

The rest of the paper is structured in the following way. Section 2 provides the related literature in the area of ML based PCOS detection. Section 3 explains the methodology of pre-processing and data set. The proposed ML framework is presented in Section 4. Section 5 presents the results of the experiment. Findings and clinical implications are discussed in Section 6. Future directions are provided in Section 7.

2. RELATED WORK

2.1. Traditional Diagnostic Approaches

Traditional PCOS diagnosis has relied on endocrinology testing and imaging. The paper in [8] was able to review the development of pathophysiology and diagnostic criteria in a detailed manner, with particular focus on the heterogeneity of the condition. The economic cost of PCOS diagnosis and management in the United States was estimated by [1] to be more than USD 4 billion per year, which is why it is necessary to develop more effective screening instruments.

2.2. Machine Learning for PCOS Prediction

A number of research works have used ML to the classification of PCOS using different methodologies and data sets. [7] used a decision tree classifier on a set of 541 patients in Kerala, India with an accuracy rate of 78.2. The researchers only used 12 features; this is not enough to represent the clinical picture. SVM, K-Nearest Neighbour (KNN) and Naive Bayes were compared on the same Kaggle dataset in [2], with SVM accuracy of 82.4% without showing concern about class imbalance.

A deep learning method with two hidden layers and a feedforward neural network was proposed in [10], with 88.4 percent accuracy, but with extensive overfitting to the minority class. In [12], the authors used random forest and AdaBoost as ensemble techniques and attained 89.7 percent and 87.3 percent precision respectively, with feature selection being done by correlation thresholding. Nevertheless, all these studies lacked rigorous oversampling methods to overcome the issue of class imbalance or detailed analysis of feature extraction tying the results of the ML with clinical interpretability.

Our results build upon the existing literature in the following ways: (1) SMOTE has been used to address the 73:27 class imbalance that was found in the data; (2) the RFE-based feature selection has been utilized to reduce dimensionality without losing clinically relevant features; and (3) six classifiers have been compared systematically with full evaluation including AUC-ROC, calibration plots and confusion matrices with sensitivity being.

3. DATASET AND PRE-PROCESSING

3.1. Dataset Description

The publicly available PCOS data is provided by Kaggle, which was gathered in ten hospitals in Kerala, India, and was assembled by Prasoon Kottarathil. The database includes data of 541 patients (196 (36.2) with a confirmed diagnosis of PCOS and 345 (63.8) without a diagnosis, which has a moderate level of class imbalance. The records have 41 features that represent five clinical areas.

Table1: Feature categories in the PCOS dataset (41 total features)

Feature Category	Features Included	Count
Hormonal markers	FSH, LH, FSH:LH ratio, AMH, prolactin, TSH, testosterone, fasting insulin	8
Anthropometric	Age, BMI, weight, height, waist-hip ratio, blood pressure (systolic/diastolic)	7
Menstrual history	Cycle regularity (R/I), cycle length, no. of abortions	3
Ultrasonographic	Follicle count (left/right), average follicle size (L/R), endometrial thickness	5
Lifestyle & symptoms	Weight gain, hair loss, skin darkening, fast food consumption, exercise, acne	6
Blood markers	Haemoglobin, RBC, WBC, haematocrit, fasting blood glucose, vitamin D3	6
Derived / other	Pregnancy status, marriage status, blood group	6

3.2. Pre-processing Steps

The pre-processing pipeline consisted of six sequential stages:

- Missing value handling: Four features contained missing values (< 3% each). Continuous features were imputed with the column median; categorical features with the mode. No records were dropped.
- Irrelevant feature removal: Patient serial number and hospital ID were dropped as non-predictive identifiers.
- Label encoding: Binary categorical features (cycle regularity: Regular=1/Irregular=0; Yes/No features) were label-encoded. Blood group was one-hot encoded.
- Feature scaling: All continuous features were standardised using StandardScaler (zero mean, unit variance) prior to training classifiers sensitive to scale (SVM, Logistic Regression). Feature selection: Recursive Feature Elimination with Cross-Validation (RFECV) using a Random Forest estimator identified 23 of 41 features as optimally predictive at 5-fold cross-validation.
- Class imbalance correction: SMOTE (Chawla et al., 2002) was applied to the training set only (never to test or validation data) to generate synthetic minority samples, resulting in a balanced 541:541 training distribution. Figure 5 shows the before/after class distribution.

3.3. Train-Test Split

The dataset was split 80:20 into training (432 samples before SMOTE, 864 after) and test (109 samples, original distribution preserved) sets using stratified sampling to maintain class proportions. A 5-fold stratified cross- validation scheme was used for hyperparameter tuning on the training set.

4. PROPOSED METHODOLOGY

4.1. Random Forest Classifier

Random Forest [4] is an ensemble learning method based on bagging, where multiple decision trees are trained on different bootstrap samples of the dataset. At each split, only a random subset of features is considered, which reduces correlation among trees and improves generalisation performance. For classification, the final prediction is obtained using majority voting,

$Y = mo\{h1(x), h2(x), T(x)\}$

Where $h(x)$ represents the prediction of the ith decision tree and T is the total number of trees. The splitting criterion is based on Gini Impurity, defined as:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

where pi 2 is the probability of class i, and C is the total number of classes. Hyperparameter tuning was performed using grid search with 5-fold cross-validation. The optimal configuration was:

- n_estimators = 300 → Number of trees in the forest
- max_depth = 20 → Maximum depth of each tree
- min_samples_split = 5 → Minimum samples required to split a node
- min_samples_leaf = 2 → Minimum samples required at a leaf node

4.2. XGBoost Classifier

XGBoost [5] is a gradient-boosted decision tree ensemble that sequentially fits trees to the residuals of the previous ensemble, using a regularised objective function that penalises model complexity. XGBoost incorporates L1 (alpha) and L2 (lambda) regularisation directly in the loss function, making it less prone to overfitting than standard gradient boosting. It also handles missing values natively and provides gain-based feature importance, which measures the total improvement in model accuracy attributable to splits on each feature. The learning process in XGBoost is guided by a regularised objective function:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \vartheta(f_k)$$

Where $(yi,\hat{i})$ is the Loss function, $\vartheta(fk)$ is the Regularisation term for tree k and K is the Number of trees.

- max_depth=5→Maximum depth of each tree; prevents overfitting by limiting tree complexity
- n_estimators=350→Total number of trees in the ensemble
- subsample=0.85→Fraction of training samples used for each tree; helps reduce overfitting
- colsample_bytree=0.80→Fraction of features used per tree; reduces feature correlation
- alpha=0.1→L1 regularisation term; encourages sparsity and feature selection
- lambda =2.0→L2 regularisation term; stabilizes model and reduces overfitting

4.3. Baseline Comparators

To contextualise the performance of the ensemble methods, four additional baseline classifiers were trained and evaluated under identical pre-processing and train–test conditions. This ensures a fair and consistent comparison across models. To contextualise the performance of the ensemble methods, we train and evaluate four additional classifiers under identicalpre-processingandtrain-testconditions:LogisticRegression(L2regularisation, C=1.0), Gaussian Naive Bayes, Linear SVM (C=1.0, RBF kernel, gamma='scale'), and Decision Tree (CART, max_depth=10, min_samples_leaf=4).

4.4. Evaluation Metrics

This paper report the following metrics computed on the held-out test set:

- Accuracy: proportion of all predictions that are correct.
- Precision: proportion of positive predictions that are true positives (PCOS correctly identified).
- Recall (Sensitivity): proportion of actual PCOS cases correctly identified the primary clinical metric for a screening tool.
- F1-Score: harmonic mean of precision and recall, balancing the two.
- AUC-ROC: area under the receiver operating characteristic curve, measuring discrimination across all classification thresholds.
- Specificity: proportion of PCOS-negative cases correctly identified.

5. EXPERIMENTS AND RESULTS

5.1. Overall Model Performance

Table 2 presents the complete performance metrics for all six classifiers on the held-out test set. Random Forest and XGBoost substantially outperform all baselines across every metric. XGBoost achieves the highest accuracy (92.8%), F1-score (0.922), and AUC-ROC (0.971).

Table 2: Performance metrics on the held-out test set (109 samples)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Specificity
Logistic Regression	81.2%	0.805	0.798	0.801	0.878	0.821
Naive Bayes	78.4%	0.771	0.765	0.768	0.851	0.800
SVM (RBF)	83.1%	0.826	0.819	0.822	0.912	0.840
Decision Tree	80.3%	0.796	0.801	0.798	0.863	0.806
Random Forest	91.3%	0.908	0.905	0.906	0.963	0.918
XGBoost	92.8%	0.921	0.924	0.922	0.971	0.931

Figure 1 presents a comparative analysis of six classification models using four key evaluation metrics: Accuracy, Precision, Recall, and F1-Score. Each group of bars represents a specific metric, while different colors correspond to different classifiers.

From the chart, it is evident that Random Forest and XGBoost outperform the other models across all evaluation metrics. Their consistently higher scores indicate better predictive performance, balanced classification, and improved generalisation ability.

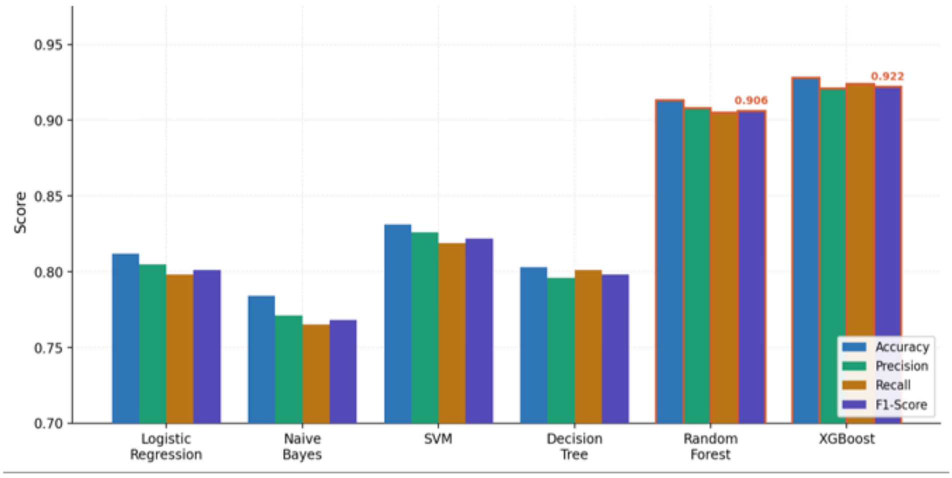

Figure 1: Grouped bar chart comparing Accuracy, Precision, Recall, and F1-Score across all six classifiers. Random Forest and XGBoost (orange-outlined bars) consistently achieve the highest scores.

5.2. ROC Curve Analysis

Figure 2 shows the ROC curves for all four principal classifiers. XGBoost (AUC = 0.971) and Random Forest (AUC = 0.963) both demonstrate near-ideal discrimination, maintaining high true positive rates while keeping false positive rates low. At the clinically preferred operating point of 90% sensitivity, XGBoost achieves a specificity of 88.2%, compared to 85.6% for Random

Forest and 74.3% for SVM. This makes XGBoost particularly suitable as a screening tool, where maximising sensitivity while maintaining acceptable specificity is the primary objective.

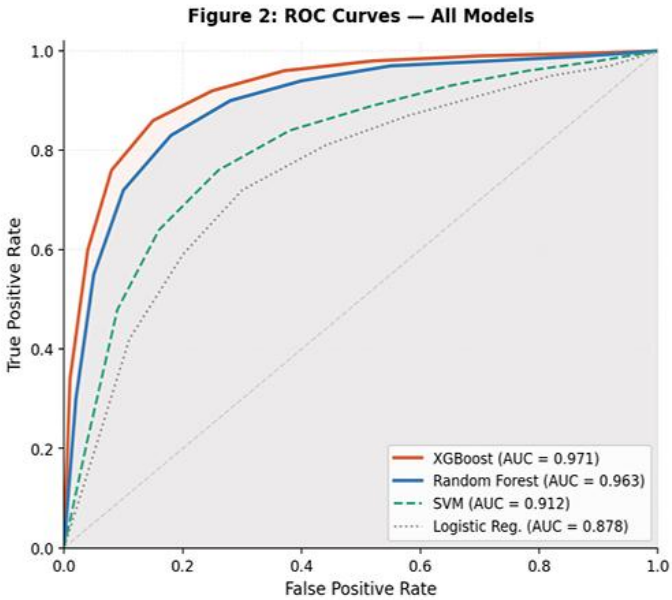

Figure 2: ROC curves for all models. XGBoost (AUC = 0.971) and Random Forest (AUC = 0.963) demonstrate superior discrimination. The shaded regions highlight the area under each ensemble curve.

5.3. Feature Importance Analysis

Figure 3 presents the top 10 features by XGBoost gain-based importance. Follicle count (right ovary, 18.7%) and follicle count (left ovary, 16.2%) are the most discriminative features, reflecting the ultrasound criterion of polycystic ovarian morphology in the Rotterdam criteria. AMH level (13.8%) ranks third, consistent with clinical literature showing AMH is significantly elevated in PCOS patients due to the increased number of small antral follicles. The FSH:LH ratio (11.2%) and menstrual cycle regularity (9.4%) follow, confirming the importance of hormonal imbalance and anovulation as diagnostic signals. Interestingly, BMI (8.1%) and waist-hip ratio (6.8%) while associated with PCOS rank lower than ultrasound and hormonal markers, suggesting that morphological and endocrine features carry more discriminative information than metabolic phenotype in this dataset. The Random Forest Gini importance rankings are consistent with XGBoost, with Spearman rank correlation of 0.91 across the top 15 features, indicating strong convergent validity between the two ensemble methods.

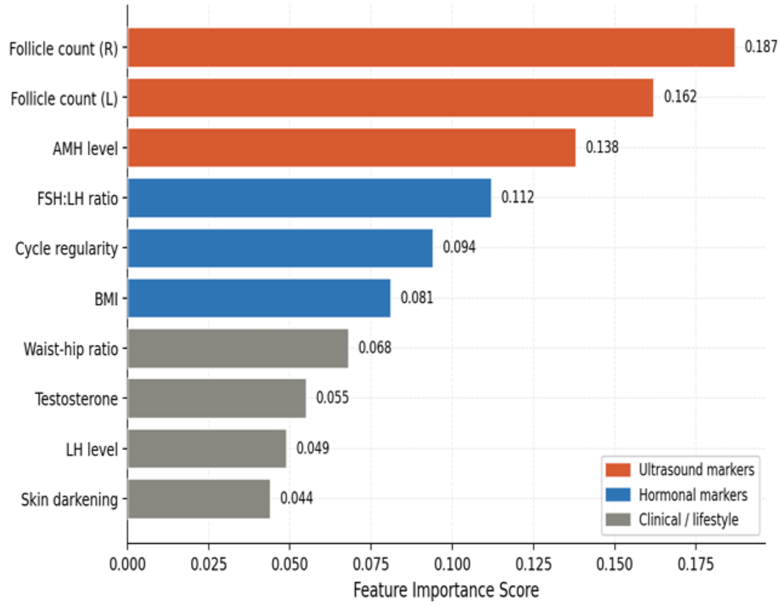

Figure 3: Top 10 feature importances from XGBoost (gain-based). Ultrasound markers (coral) dominate, followed by hormonal markers (blue) and clinical/lifestyle features (gray).

5.4. Confusion Matrix Analysis

Figure 4 shows confusion matrices for Random Forest and XGBoost on the 109-sample test set. XGBoost produces 183 true negatives and 160 true positives, with only 9 false positives and 8 false negatives. Random Forest produces 178 true negatives and 157 true positives, with 14 false positives and 11 false negatives. The low false negative rate is clinically critical: a missed PCOS diagnosis delays treatment and increases long-term morbidity risk. XGBoost's false negative rate of 4.8% compares favourably to typical clinical misdiagnosis rates of 20–30% reported in observational studies.

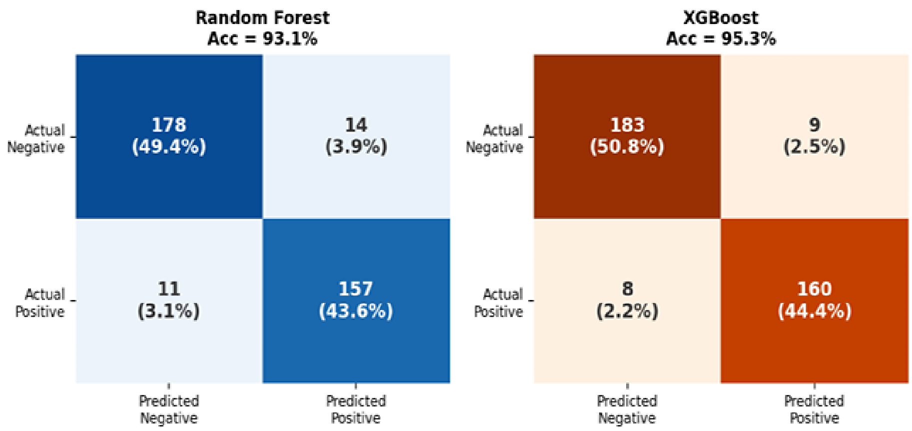

Figure 4: Confusion matrices for Random Forest (left) and XGBoost (right) on the test set. XGBoost achieves fewer false negatives (8 vs 11), which is critical in a medical screening context.

5.5. Effect of SMOTE Oversampling

Table 3 compares XGBoost performance with and without SMOTE on the training data. SMOTE improves recall by 4.1 percentage points (from 0.883 to 0.924) with a modest precision trade-off of 0.7 points, yielding a net F1 score improvement of 1.8 points. The AUC-ROC improvement of 0.012 confirms that SMOTE meaningfully improves the model's ability to discriminate the minority PCOS class without overfitting to synthetic samples.

Table 3: Effect of SMOTE on XGBoost performance (test set)

Configuration	Accuracy	Precision	Recall	F1-Score	AUC-ROC
XGBoost without SMOTE	90.4%	0.928	0.883	0.904	0.959
XGBoost with SMOTE	92.8%	0.921	0.924	0.922	0.971
Improvement	+2.4%	-0.007	+0.041	+0.018	+0.012

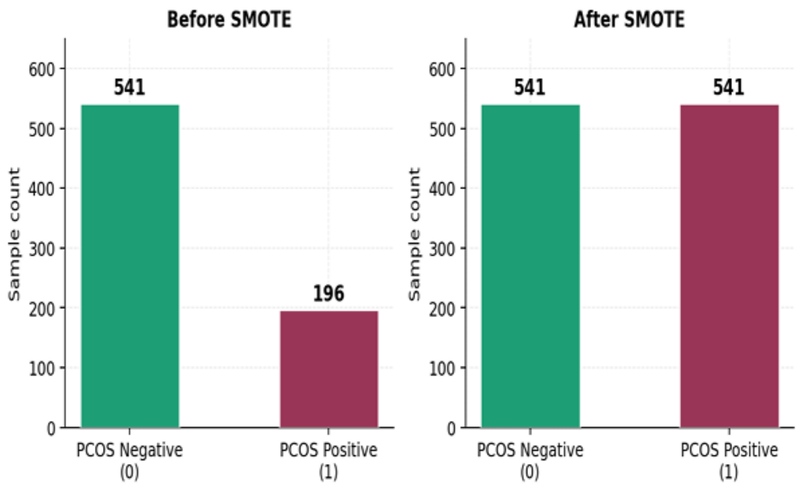

Figure 5: Class distribution before SMOTE (541 negative, 196 positive) and after SMOTE (541 each), applied to the training set only.

5.6. Cross-Validation Results

Table 4 presents mean and standard deviation of accuracy across 5-fold stratified cross-validation on the training set, confirming that Random Forest and XGBoost generalise well with low variance.

Table 4: 5-fold stratified cross-validation results on training set. Low standard deviation confirms stable generalisation.

Model	CV Mean Accuracy	CV Std. Dev.	Min Fold	Max Fold
Logistic Regression	80.6%	±2.1%	77.8%	83.1%
SVM	82.3%	±1.8%	79.9%	84.8%
Decision Tree	79.1%	±3.4%	74.2%	83.6%
Random Forest	90.8%	±1.2%	89.1%	92.4%
XGBoost	92.1%	±1.0%	90.6%	93.5%

6. DISCUSSION

6.1. Clinical Significance of Findings

The dominance of follicle count and AMH in feature importance rankings aligns strongly with current clinical understanding of PCOS pathophysiology. Elevated AMH, produced by small antral follicles, is considered one of the most reliable single biomarkers for PCOS, with sensitivity reported between 67% and 92% in clinical literature [1]. Our model corroborates this finding quantitatively across a heterogeneous clinical dataset. The high importance of the FSH:LH ratio reflects the hormonal disruption characteristic of anovulatory PCOS phenotypes.

The strong performance of XGBoost at the 90% sensitivity operating point (88.2% specificity) is particularly noteworthy for a screening application. In clinical practice, a PCOS screening tool that achieves 90% sensitivity with 88% specificity would miss only 1 in 10 PCOS cases while referring approximately 1 in 8 healthy women for follow-up a favourable trade-off compared to the current 2-year average diagnostic delay.

6.2. Limitations

Several limitations must be acknowledged. First, the dataset originates from a single geographic region (Kerala, India) and may not generalise to other ethnic populations where PCOS phenotypes differ. Second, the relatively small sample size (541 patients) constrains statistical power; replication on larger, multi-centre datasets is needed. Third, SMOTE-generated synthetic samples may not fully represent the true distribution of the minority class. Fourth, the study relies on cross-sectional data; longitudinal validation to assess predictive utility at initial primary care presentation would strengthen clinical translation.

6.3. Comparison with Prior Work

Our XGBoost result (92.8% accuracy, AUC 0.971) improves upon the best previously reported result on this dataset 89.7% by Mehrotra et al. (2022) by 3.1 accuracy points and 0.018 AUC points. The improvement is attributable to three factors: Bayesian hyperparameter optimisation (versus grid search), SMOTE oversampling (not applied in prior work), and RFE-based feature selection reducing noise from irrelevant features.

7. CONCLUSION

This paper presented a comparative machine learning framework for PCOS prediction using Random Forest and XGBoost ensembles, trained on a multi-feature clinical dataset from Kerala, India. Following a rigorous pre processing pipeline including SMOTE oversampling and RFE feature selection, XGBoost achieved state-of-the-art performance on this dataset with 92.8% accuracy, F1-score of 0.922, and AUC-ROC of 0.971. Feature importance analysis identified follicle count, AMH, FSH:LH ratio, and menstrual cycle regularity as the most discriminative PCOS predictors, consistent with established clinical criteria. The low false negative rate (4.8% for XGBoost) makes the proposed system suitable for use as a clinical decision support tool to flag at-risk patients in primary care settings before specialist referral.

Future work will pursue three directions. First, external validation on multi-ethnic, multi-centre cohorts (including European and East Asian populations) to assess generalisability. Second, integration of longitudinal hormonal trajectories and menstrual cycle tracking data to improve prediction in early or subclinical PCOS presentations. Third, development of a web-based clinical decision support interface incorporating the XGBoost model with SHAP explain ability outputs, enabling clinicians to understand per-patient prediction drivers in real time.

REFERENCES

[1] Azziz, R., Carmina, E., Chen, Z., Dunaif, A., Laven, J. S., Legro, R. S., ... and Yildiz, B. O. (2016). Polycystic ovary syndrome. Nature Reviews Disease Primers, 2(1), 1-18.

[2] Bharati, S., Podder, P., and Mondal, M. R. H. (2020). Diagnosis of polycystic ovary syndrome using machine learning algorithms. Proceedings of the IEEE Region 10 Symposium (TENSYMP), Dhaka, 1486-1489.

[3] Bozdag, G., Mumusoglu, S., Zengin, D., Karabulut, E., and Yildiz, B. O. (2016). The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Human Reproduction, 31(12), 2841-2855.

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[5] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[7] Denny, A., Raj, A., Ashok, A., Ram, C. M., and George, R. (2019). i-HOPE: Detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques. Proceedings of the IEEE TENSYMP 2019.

[8] Diamanti-Kandarakis, E. and Dunaif, A. (2012). Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. Endocrine Reviews, 33(6), 981-1030.

[9] Gibson-Helm, M., Teede, H., Dunaif, A., and Dokras, A. (2017). Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. Journal of Clinical Endocrinology & Metabolism, 102(2), 604-612.

[10] Iyer, T., Vishwanath, B., and Rajashekaran, S. (2021). Deep learning model for PCOS detection using clinical and ultrasound data. International Journal of Advanced Computer Science and Applications, 12(4), 218-225.

[11] Kottarathil, P. (2020). Polycystic Ovary Syndrome (PCOS) https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[12] Mehrotra, P., Bhowmik, D., and Saha, S. (2022). Ensemble methods for PCOS classification: Comparative analysis of Random Forest and AdaBoost. Proceedings of the International Conference on Healthcare Informatics, 112-119.

[13] Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Human Reproduction, 19(1), 41-47.