International Journal of Basic and Applied Sciences, 14 (8) (2025) 97-105

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/36w9s378
Research paper

Enhanced Diagnosis of Polycystic Ovarian Syndrome Using Stacking Classifier with Random Forest As a Meta Learner

Pragati Patil [1] *, Nandini Chaudhari [2]

[1] Research Scholar, Drs. Kiran and Pallavi Patel Global University, Vadodara, INDIA
[2] Director, Krishna School of Emerging Technology and Applied Research, Drs. Kiran and Pallavi Patel Global University, Vadodara, INDIA
**Corresponding author E-mail: phdscholar21010@kpgu.ac.in*

Received: October 28, 2025, Accepted: November 27, 2025, Published: December 4, 2025

Abstract

Polycystic Ovary Syndrome (PCOS) is considered a very serious health problem among the women of childbirth age. Early and accurate detection is essential for effective management but remains challenging due to symptom variability and subjective traditional diagnostics. Machine learning (ML) algorithms have recently emerged as promising tools for PCOS diagnosis by analyzing multidimensional clinical, hormonal, and imaging data. This study evaluates and compares multiple supervised ML methods—including Gradient Boosting (XGBoost, LightGBM, and CatBoost), AdaBoost, Logistic Regression, Naive Bayes, Random Forest, Decision Tree, Support Vector Machine (SVM), and Stacking classifiers—on a publicly available dataset of 541 patients. Various evaluation measures like Receiver Operating Characteristic curve (AUC-ROC), recall, accuracy, F1-score, and precision are used for the evaluation. Stacking classifier outperforms all single models used in this research. The Stacking classifier with Random Forest as meta-learner achieved 98% accuracy with AUC of 98%. These findings demonstrate that advanced ensemble ML models can robustly detect PCOS, with potential for integration into clinical workflows for early, objective diagnosis. Future work should focus on multi-center validation and explainable AI to enhance clinical trust and deployment.

Keywords: *Polycystic Ovary Syndrome (PCOS); Machine Learning; Ensemble Methods; Feature Selection; Diagnostic Accuracy.*

1. Introduction

Polycystic ovary syndrome (PCOS) is considered a very serious health problem among the women of childbirth age. The very common signs of PCOS are irregular mensuration, hair increasing in the body, weight gaining, and acne. It can increase the possibility of developing type 2 and gestational diabetes. So, it is necessary to diagnose it very early for an effective treatment. It is very crucial to develop some methods for helping the doctors in diagnosing PCOS very early. [1 - 3].

Recent developments in artificial intelligence techniques like machine learning (ML) facilitate objective, data-driven analyses of large, heterogeneous clinical datasets, enabling improved PCOS detection [4]. Supervised ML algorithms, comprising Support Vector Machine (SVM), Gradient Boosting Machines (GBM), Random Forest (RF), and neural networks, have shown promising classification accuracy [5] [6]. Ensemble and stacking methods combining multiple classifiers often yield superior results by exploiting complementary strengths of individual models [7]. Challenges remain, however, in feature selection, balancing classes, model generalization across diverse populations, and clinical interpretability.

This study employed diverse supervised ML methods for the diagnosis of PCOS. A 541 patient records dataset available on Kaggle is used in this study to evaluate and compare the diagnostic performance of classical, ensemble, boosting, and meta-ensemble methods, considering the strengths and limitations of each approach. The methodology leverages advanced model tuning, balanced training procedures, and interpretability measures to maximize clinical relevance.

The approach has used rigorous preprocessing, feature engineering, and hyperparameter tuning. The goals are to (i) compare the diagnostic performance of classical and ensemble classifiers, (ii) identify critical predictive features, and (iii) discuss translational implications for clinical use.

Khanna et al. (2023) developed a metalearner based stacking framework using various classifiers (RF, SVM, KNN, AdaBoost, Extra Trees) on a clinical CSV dataset of 541 samples. They implemented feature engineered inputs and evaluated model interpretability using SHAP, LIME, ELI5, and Qlattice to increase transparency. Their ensemble achieved 98% accuracy, outperforming DNN and 1D CNN baselines. They emphasized potential real time deployment and validated the approach as more meaningful for clinicians [8]. Elmannai et al. (2023) used optimized feature selection and explainable AI on a similar 541 sample dataset. They applied SMOTE to address class imbalance, Bayesian optimization for hyperparameters, and stacking combinations of tree based classifiers. This approach achieved nearly perfect detection accuracy (100%), with interpretability ensured via feature importance explainability tools [9]. Panjwani et al. (2025) introduced an optimized ML pipeline for early PCOS detection. They used feature selection, scaling, hyperparameter tuning, and ensemble classifiers including RF, SVM, LR, KNN, and LGBM. The final model achieved 99.8% accuracy (LR/SVM) on clinical data, highlighting importance of hybrid feature selection strategies and extensive validation [10]. Hossein et al. (2022) proposed PCONet, a custom CNN architecture trained on ovarian ultrasound images to detect PCOS. Compared to fine-tuned InceptionV3, PCONet achieved 98.12% accuracy vs. 96.56%, demonstrating image based ML's strength. They emphasized architectural simplicity and efficiency for medical imaging tasks [11]. Lv et al. (2022) explored deep learning on bodily images—not traditional tabular data—using scleral (eye) images from 721 women (388 PCOS cases). Employing U Net segmentation with ResNet feature extraction and multi instance learning, they achieved AUC = 0.979 and accuracy = 0.929, pointing to novel noninvasive sensing approaches in PCOS screening [12].

Mohi Uddin et al. (2025) presented a comparative analysis, reviewing prior studies (Silva, Rahman, Khanna, Bharati, Zigarelli, and Danaei Mehr) and highlighting their constraints. They then proposed their own pipeline using mutual information for feature selection and ensemble models (RF, AdaBoost), achieving accuracy up to 99.78% on 541 sample datasets, improving generalizability over earlier work [13]. Rahman et al. (2024) implemented a web based diagnostic system using LR, DT, AdaBoost, and RF on the same 541 record dataset. They reported ensemble methods (especially AdaBoost + RF) yielding 94% accuracy, and discussed limitations like small dataset size and missing external validation [14]. Bharati et al. (2022) tested a hybrid RFLR (Random Forest + Logistic Regression) combined with univariate feature selection on 541 patients (177 PCOS positive). They used hold out and cross validation splits and reached 91% accuracy. Their analysis highlighted potential limitations of univariate feature selection ignoring interactions [15].

Zigarelli et al. (2022) applied CatBoost classifier on a Kaggle dataset of 541 individuals with 44 features. Using K fold validation, they differentiated invasive and noninvasive self-diagnosis contexts; the noninvasive clinical indication model achieved 90.1% accuracy. They discussed ease of self-screening but noted accuracy gaps in invasive treatment prediction [16]. Danaei Mehr & Polat (2024) examined ensemble combining Extra Trees, AdaBoost and RF, and MLP in tabular PCOS diagnosis. Their best ensemble RF model reached 98.89% accuracy using reduced feature subsets, indicating that smaller informative feature sets can yield very high performance [17]. Zad et al. (2024) conducted modeling on a large EHR dataset (30,601 records), using RF, GBT, SVM, LR to predict PCOS ICD diagnoses prior to clinical recognition. Gradient boosted trees achieved best AUC (85%), but accuracy/AUC lower than small datasets, revealing challenges with scale and class heterogeneity [18].

Elmannai et al. (2023) emphasized per feature explainability using heat maps and saliency techniques along with SMOTE and Bayesian optimization; they benchmarked their model extensively and discussed scaling to real time mobile applications [19]. Divekar & Sonawane (2024) submitted a study "AUTO PCOS" exploring transfer learning (InceptionV3) on ultrasound frames, achieving 90.52% accuracy and >90% precision/recall/F1, with interpretability via LIME and saliency maps, illustrating ultrasound based diagnostic pipelines on new datasets [20]. Morris et al. (2023) proposed federated learning to preserve patient privacy while modeling PCOS treatment optimization. Using synthetic PCOS patient data, they demonstrated that FL can generalize across institutions without sharing raw data, a promising route for multicenter collaboration [21]. Multiple small scale studies (2022–2024) like Silva et al. (RF with BorutaShap, 93% accuracy on 72/73 subjects) also contributed to methodological foundations, although mostly lacking external validation and dataset diversity [22].

The paper is structured as follows. The dataset, preprocessing steps, and methodology and models used are given in section 2. Experiments and results are discussed in section 3. Finally we conclude with the future work directions in section 4.

2. Methodology

2.1. Dataset

A freely accessible dataset from Kaggle is used in this study, containing records of 541 patients with 41 demographic, clinical, hormonal, biochemical, and ultrasound attributes relevant to PCOS diagnosis [23][08]. The dataset includes 364 PCOS-negative and 177 PCOS-positive samples, reflecting a moderate class imbalance, and initial inspection revealed missing values. Although widely used in PCOS research, this dataset has important limitations. Because it originates from a single clinical source, it may not fully represent broader demographic diversity (e.g., ethnicity, geographic region, age distribution), introducing potential demographic bias. Similarly, clinical variability is constrained by uniform data collection procedures and diagnostic criteria at the source institution. As real-world clinical environments differ in laboratory ranges, imaging interpretation, and diagnostic thresholds, model performance may vary when applied to external populations.

Nevertheless, the dataset remains valuable as one of the few publicly available, comprehensive PCOS datasets, enabling reproducible methodological development. Thus, the findings of this study should be viewed as proof-of-concept rather than fully generalizable. To address data quality issues, median imputation was applied to missing numerical features and mode imputation to categorical variables. Categorical attributes (e.g., blood group, symptom indicators) were encoded using label encoding and one-hot encoding. Continuous features with differing scales were standardized using z-score normalization. To mitigate class imbalance, SMOTE was applied to the training set to synthetically increase PCOS-positive samples. Feature selection was performed using Recursive Feature Elimination (RFE) and mutual information to retain the most predictive attributes and reduce noise. The dataset was split using an 80:20 stratified train–test ratio to preserve class distribution. This preprocessing pipeline ensured a balanced, high-quality dataset suitable for robust machine-learning modeling.

2.2. Machine learning models

This study applies a comprehensive set of supervised machine learning (ML) algorithms to classify Polycystic Ovary Syndrome (PCOS) status using a carefully preprocessed clinical dataset.

Logistic Regression models the log-odds of the probability of PCOS presence as a linear combination of input features. It is commonly employed for binary classification given its simplicity, interpretability, and efficiency. LR uses maximum likelihood estimation to fit coefficients and outputs probabilistic predictions, which can inform clinical risk assessments. Regularization techniques, such as L2 (Ridge) or L1 (Lasso), are utilized to reduce overfitting, especially in high-dimensional feature spaces [24].

Decision Trees recursively partition the dataset by selecting features and threshold splits that maximize class purity (e.g., via Gini impurity or entropy). While easy to interpret, single DTs are prone to overfitting and sensitive to data noise. To enhance robustness, Random Forest aggregates predictions from hundreds of DTs trained on bootstrapped samples and random feature subsets. This ensemble reduces variance and bias, improving generalization. RF models also output feature importance scores, guiding insights into key clinical and biochemical predictors of PCOS [25]

SVM algorithm finds a hyperplane which separates hyperplane in such a way that it increases the margin among classes in a transformed feature space, often using kernels such as Radial Basis Function (RBF) to capture nonlinear relationships. SVMs work well with high-dimensional data and limited samples, both common in medical datasets. Hyperparameters include the penalty parameter C controlling regularization strength and kernel parameters (e.g., gamma for RBF). Grid or Bayesian searches optimize these to balance bias-variance trade-offs [26].

AdaBoost iteratively trains weak learners—generally shallow decision trees—on weighted data samples, emphasizing previously misclassified observations. These weak learners combine adaptively into a strong classifier via weighted majority voting. AdaBoost improves accuracy and noise robustness in medical datasets characterized by heterogeneous patterns. Parameters include the number of estimators and learning rate controlling model complexity [24].

Gradient Boosting models build additive learners sequentially by fitting new models to residual errors of prior models, optimizing a differentiable loss function. Key variants used include:

- XGBoost: Incorporates regularization, tree pruning, and parallelization improvements, widely adopted in clinical predictive modeling for its superior accuracy [24].
- LightGBM: Utilizes histogram binning and leaf-wise tree growth strategies for enhanced computational speed and lower memory usage, efficiently handling categorical variables common in clinical data [24].
- CatBoost: Addresses prediction bias stemming from categorical data through ordered boosting, natively handling categorical features without extensive preprocessing, making it robust in heterogeneous medical datasets [25].

2.3. Hyperparameter tuning

Hyperparameter tuning for boosting models includes controlling tree depth, learning rates, number of estimators, and regularization coefficients to minimize overfitting and optimize predictive performance.

Naive Bayes classifiers apply Bayes' theorem assuming conditional independence among features. Despite this strong assumption, NB can perform competitively in medical classification due to its simplicity, robustness with small datasets, and fast training. NB outputs posterior probabilities, aiding risk stratification [27].

2.4. Stacking

Stacking is an advanced ensemble technique combining multiple disparate base learners into a meta-classifier that aggregates their predictions to improve overall accuracy and reduce individual model biases. In this work, base classifiers include LR, DT, RF, SVM, AdaBoost, GBM variants (XGBoost, LightGBM, CatBoost), and NB. The meta-classifier is selected from among Random Forest or Logistic Regression. The model is trained on base learners' out-of-fold predictions to avoid data leakage, thereby increasing robustness and leveraging complementary decision boundaries from diverse classifiers [24].

Figure 1 shows the methodology adopted in this work.

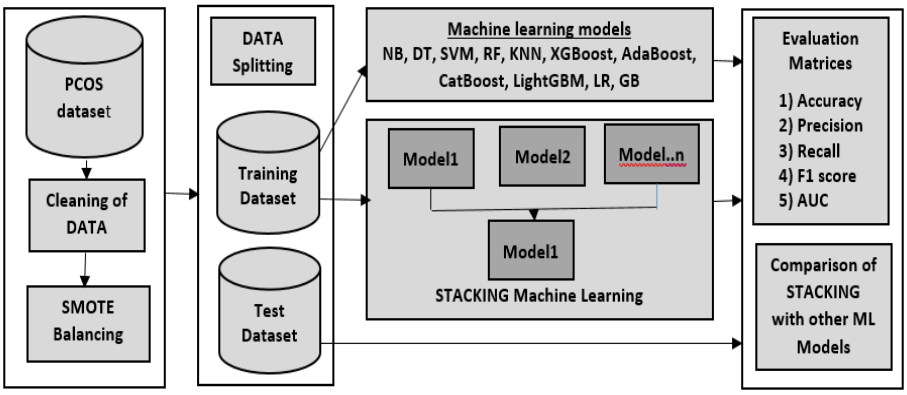


Fig. 1: Methodology.

2.5. Model training and evaluation procedure

- Data Splitting: Following pre-processing, we have split the dataset in a ratio of 80 (training):20(test). This helped us in preserve class distribution [27].
- Class Imbalance Handling: The minority PCOS-positive class is oversampled synthetically via Synthetic Minority Over-sampling Technique (SMOTE) on the training data, mitigating bias toward the classes and improving sensitivity [27].
- Hyperparameter Optimization: Grid search and Bayesian optimization frameworks are applied across classifiers to identify optimal hyperparameters such as tree depths, learning rates, number of estimators, SVM regularization, and kernel parameters. Early stopping criteria prevent overfitting during models' training [24].
- Feature Selection: Recursive Feature Elimination (RFE) and impurity-based feature importance from RF and boosting models guide dimensionality reduction, improving interpretability and reducing computation without sacrificing accuracy [27].
- Evaluation Metrics: Multiple metrics as mentioned above are used for the evaluation. Values of the metrics such as precision, accuracy, F1-score, recall, and AUC-ROC—are calculated to comprehensively assess classifier performance, emphasizing balanced metrics given the medical context's importance of detecting true positives [24].

This detailed methodology thus combines thorough data handling, diverse ML models from interpretable regressors to complex ensembles, and rigorous evaluation to identify the most effective PCOS detection models for potential clinical integration.

3. Results and Discussion

The performance of various machine learning (ML) algorithms for PCOS detection was evaluated using multiple quantitative metrics and visualization tools to comprehensively assess classification effectiveness. The metrics utilized were accuracy, precision, recall (sensitivity), F1-score, and Area under the Receiver Operating Characteristic Curve (AUC-ROC).

Accuracy Measures the proportion of total correct predictions (both positive and negative) over all samples. Higher accuracy indicates better overall performance. Precision is the ratio of true positive predictions to all positive predictions, reflecting the classifier's ability to avoid false positives. Recall (Sensitivity) is the proportion of true positive cases correctly identified, critical in clinical settings to minimize missed diagnoses. F1-Score is the harmonic mean of precision and recall, providing a balanced metric especially relevant for imbalanced datasets. AUC-ROC Measures the model's ability to distinguish between classes across different thresholds; values closer to 1 indicate excellent discriminative capability.

Additional qualitative analysis was performed through ROC curves, confusion matrices, and detailed classification reports. The stacking classifier employed Random Forest as the meta-learner, and among the boosting algorithms, LightGBM and CatBoost were included to leverage their advanced gradient boosting frameworks.

Table 1: Performance Achieved by the Models Used in This Work

Models	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.95	0.95	0.95	0.95	0.98
Naive Bayes	0.91	0.86	0.99	0.92	0.98
XGBoost	0.95	0.96	0.95	0.95	0.99
AdaBoost	0.95	0.93	0.96	0.95	0.98
Gradient Boosting	0.96	0.96	0.96	0.96	0.98
Logistic Regression	0.93	0.9	0.97	0.93	0.98
Support Vector Machine	0.54	0.59	0.26	0.36	0.55
K-Nearest Neighbors	0.68	0.65	0.81	0.72	0.75
Decision Tree	0.92	0.91	0.93	0.92	0.92
Stacking Classifier	0.98	0.99	0.93	0.96	0.98
LightGBM	0.97	0.97	0.96	0.97	0.98
CatBoost	0.95	0.95	0.95	0.95	0.98

The tabulated results for the models are displayed in TABLE 1. As seen, the stacking classifier outperforms rest of the models by achieving 98% accuracy and AUC-ROC of 98% also. This confirms the advantage of combining multiple base learners to capture complementary patterns in the data. LightGBM and CatBoost also performed strongly among individual models, leveraging their gradient boosting enhancements suited for tabular medical data. The confusion matrix for the stacking classifier highlights its robust classification capability. The high True Positive and True Negative counts imply excellent sensitivity and specificity, vital for clinical decision support. ROC curves are plotted for all models and the stacking classifier demonstrates superior curve profiles. Detailed classification reports provided precision, recall, and F1-scores per class, confirming balanced performance across minority (PCOS-positive) and majority classes, crucial to avoid bias and ensure reliable diagnosis. Figure 2 shows the comparison of the models used across all evaluation metrics.

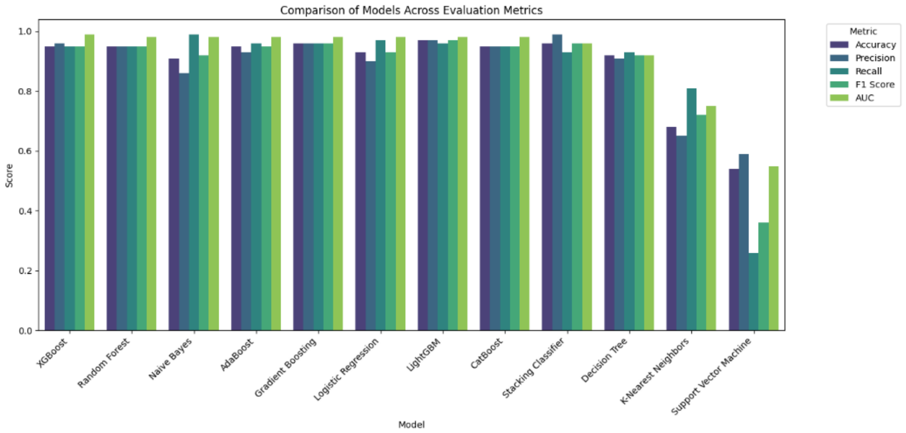
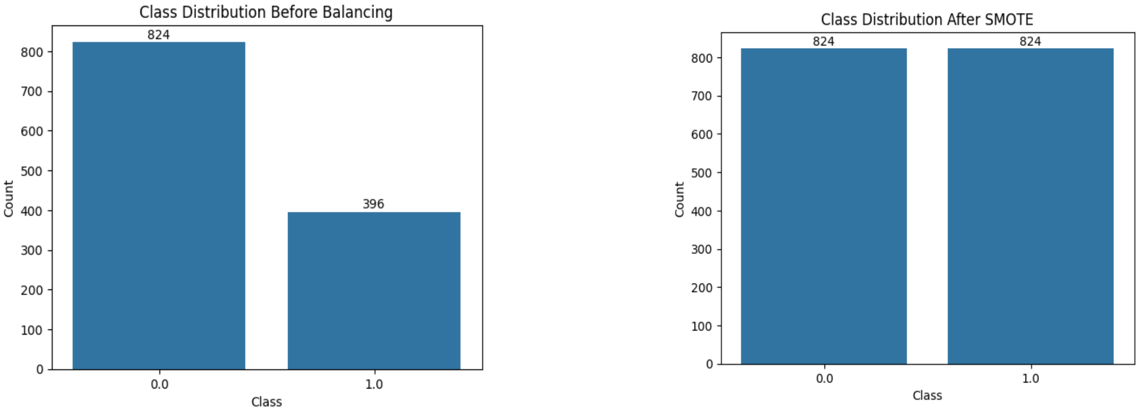


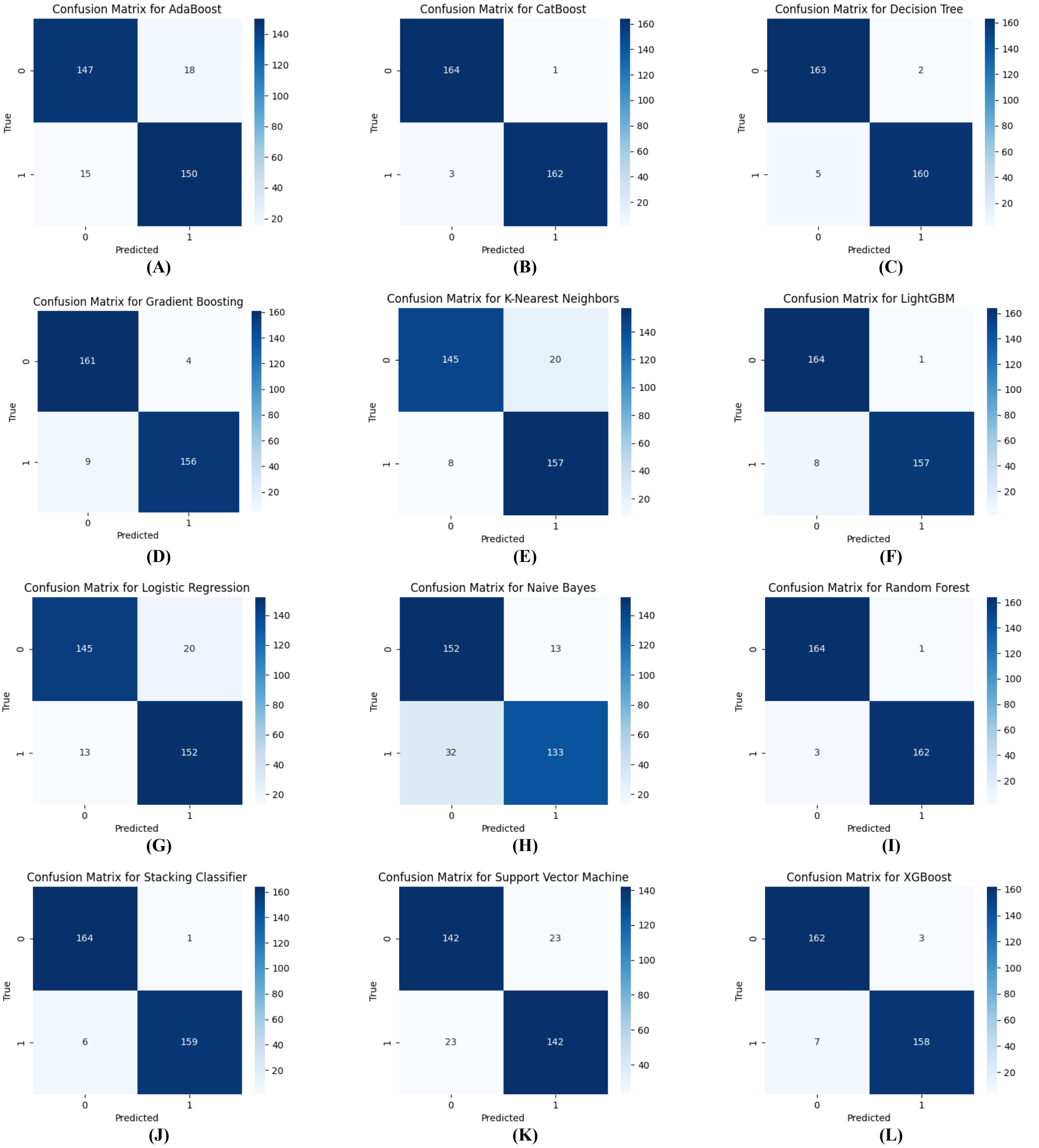
Fig. 2: Comparison of the Models Across Evaluation Metrics.

These results collectively confirm that ensemble-based methods, particularly stacking with Random Forest as the meta-classifier, combined with advanced boosting algorithms LightGBM and CatBoost, provide significant improvements in PCOS detection accuracy and reliability over traditional single classifiers. Such models are well-suited for integration into clinical workflows for early and objective PCOS diagnosis. Figure 3 shows the class distribution before and after balancing using SMOTE. Figure 4 shows the confusion matrices achieved for all models used. Figure 5 and 6 shows loss and accuracy graphs for the all models used. Figure 7 shows ROC curves for all applied models. Table 2 shows the performance comparison of proposed approach with existing approaches on the basis of Accuracy.



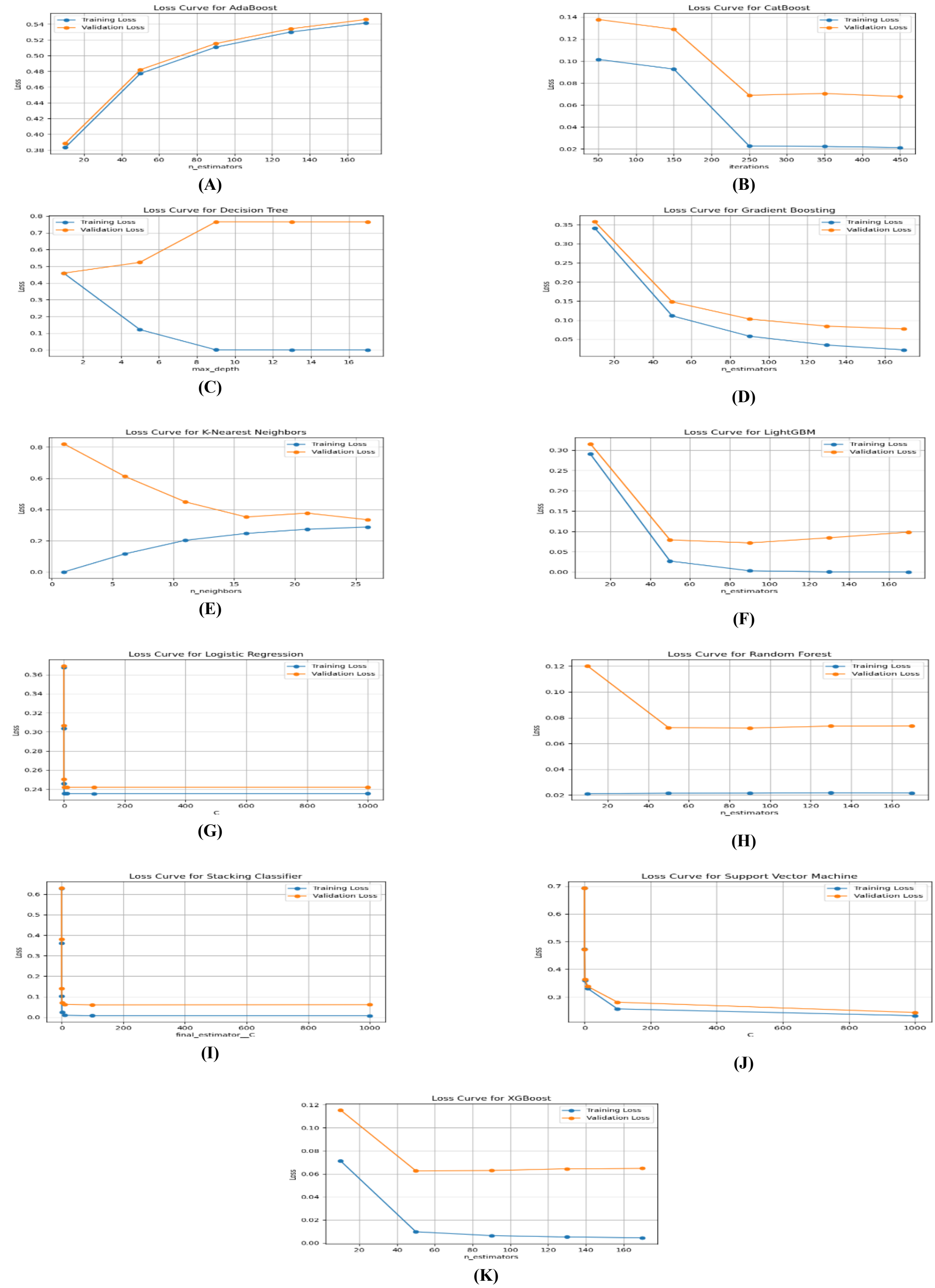
(A) (B)

Fig. 3. Class Distribution: (A) Before SMOTE (B) After SMOTE.



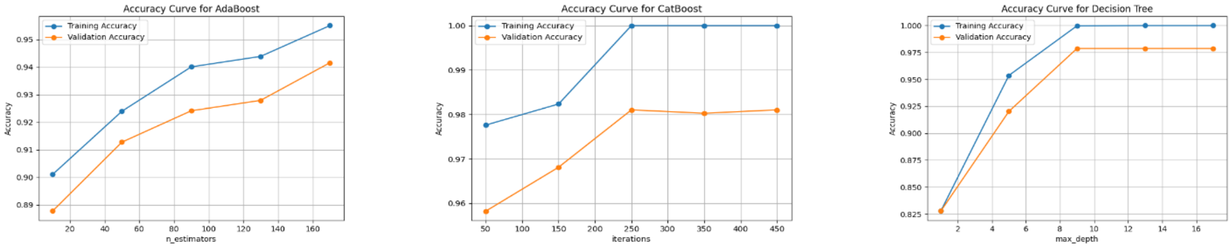
(A) (B) (C) (D) (E) (F) (G) (H) (I) (J) (K) (L)

Fig. 4. Confusion Matrix: (A) Adaboost (B) Catboost (C) Decision Tree (D) Gradient Boosting (E) K Nearest Neighbors (F) Lightgbm (G) Logistic Regression (H) Naïve Bayes (I) Random Forest (J) Staking (K) Support Vector Machine (L)Xgboost.



(A) (B) (C) (D) (E) (F) (G) (H) (I) (J) (K)

Fig. 5. Loss Curves (A) Adaboost (B) CatBoost (C) Decision Tree (D) Gradient Boosting (E) K Nearest Neighbors (F) LightGBM (G) Logistic Regression (H) Random Forest (I) Staking (J) Support Vector Machine (K) XGBoost.



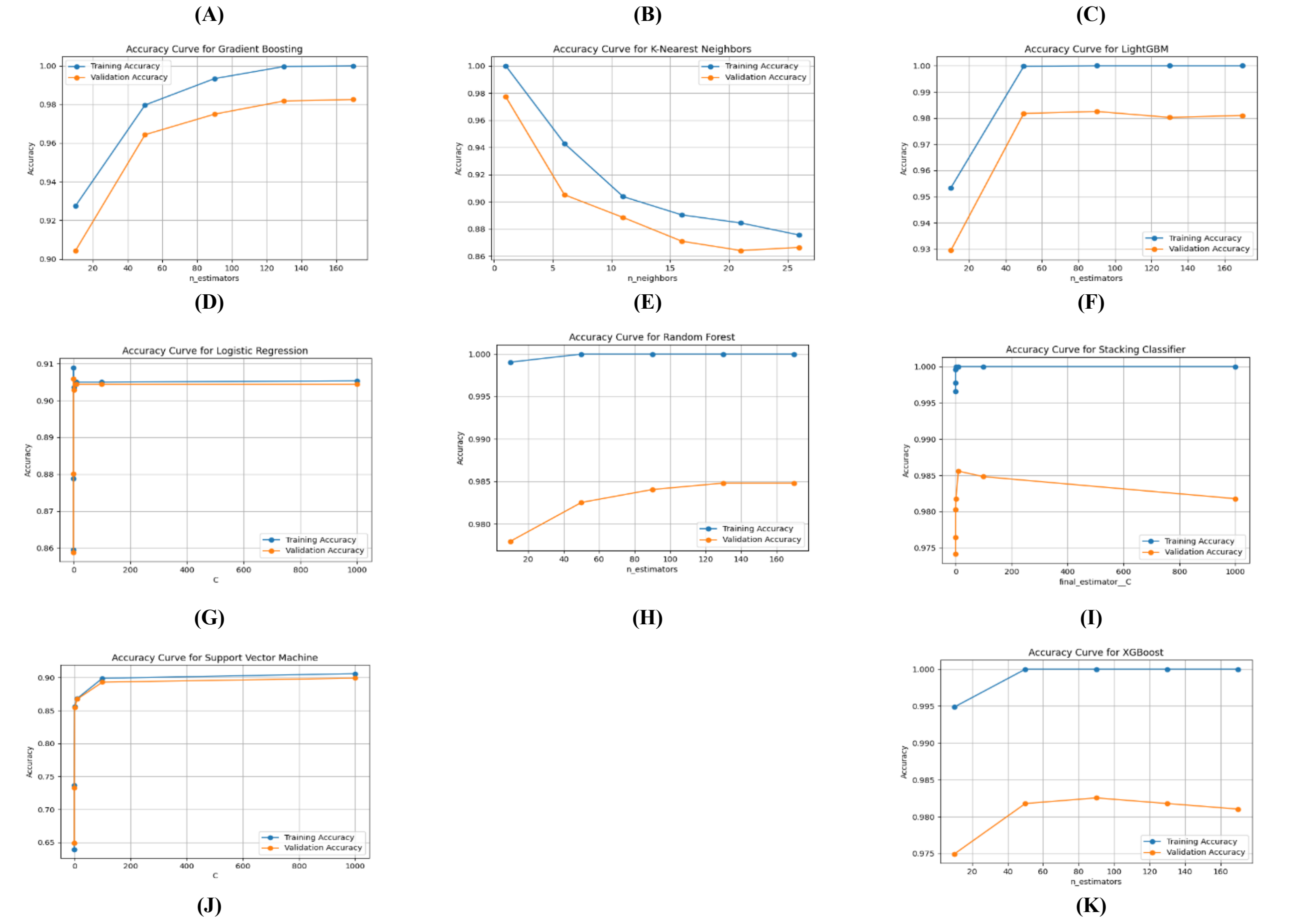


Fig. 6. Accuracy Curves: (A) Adaboost (B) CatBoost (C) Decision Tree (D) Gradient Boosting (E) K Nearest Neighbors (F) LightGBM (G) Logistic Regression (H) Random Forest (I) Staking (J) Support Vector Machine (K) XGBoost.

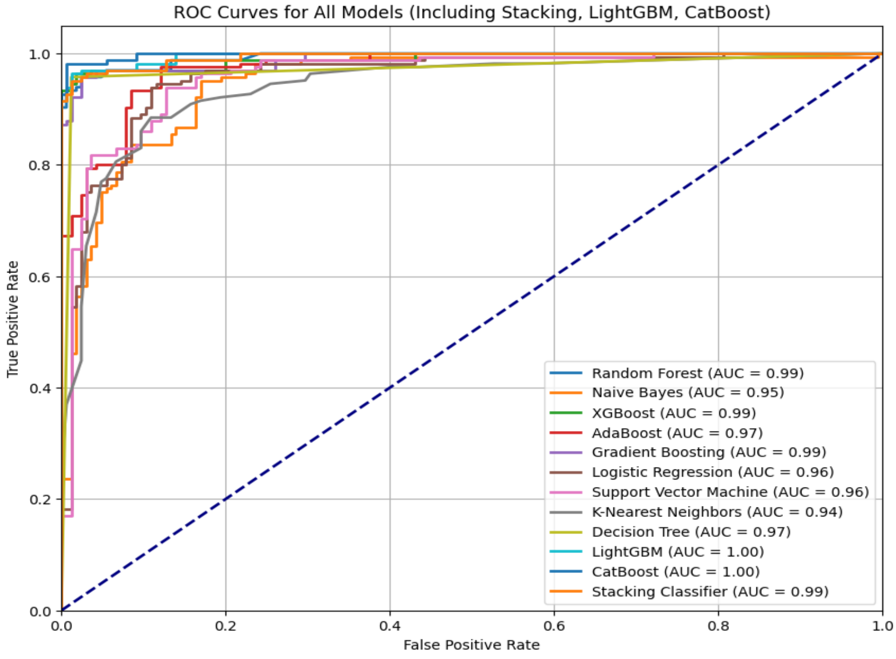


Fig. 7. ROC Curves (All Models).

Table 2. Performance Comparison of Proposed Approach with Existing Approaches on Basis of Accuracy LR: Logistic Regression, DT: Decision, Tree, RF: Random Forest, SVM: Support Vector Machine, ADB: Adaboost, GB: Gradient Boosting, LGBM: LightGBM, Catb: CatBoost, NB: Naive Bayes, STRF: Stacking Classifier with Random Forest, and KNN: K-Nearest Neighbor

Author	Models											
	RF	NB	XGB	ADB	GB	LR	SVM	KNN	DT	STRF	LGBM	CatB
[08]	0.93	0.83	0.95	0.95	-	0.89	0.93	0.90	0.87	0.93	-	-
[27]	96.83	0.95	0.96	0.96	-	0.92	0.96	0.93	0.93	0.97	-	-
[28]	0.91	0.87	0.89	0.90	0.93	0.92	0.93	0.86	0.86	0.92	-	0.91
[29]	0.92	-	0.91	-	-	0.88	-	0.76	-	0.92	-	-

[30]	0.89	-	-	0.85	-	0.81	0.84	0.70	0.83	-	0.89	0.89
[31]	0.90	0.83	0.86	-	-	-	0.67	0.67	-	-	-	-
[32]	0.69	-	-	-	-	0.72	0.72	0.69	0.71	-	-	-
[33]	0.86	0.70	0.86	0.83	0.89	0.76	0.87	0.86	0.80	0.86	-	-

To assess whether the observed differences in classification accuracy across models were statistically significant, we conducted pairwise paired t-tests using the five-fold cross-validation accuracies for each model. Since the same data folds were used across all models, a paired testing strategy is statistically appropriate. For each model pair, we computed the paired t-statistic (df = 4), two-sided p-value, and Cohen's d as an effect-size measure for the paired differences. In total, 66 model pairs were evaluated. At the conventional significance level of $p < 0.05$ (uncorrected), 54 out of 66 comparisons showed statistically significant differences. To control for family-wise error due to multiple comparisons, a Bonferroni correction was applied, resulting in a corrected significance threshold of $\alpha = 0.05/66 \approx 0.00076$, under which 35 out of 66 comparisons remained statistically significant.

The results confirm that the Stacking Classifier (mean accuracy ≈ 0.98) and LightGBM (mean accuracy ≈ 0.97) significantly outperform weaker models such as Support Vector Machine (mean accuracy ≈ 0.54) and K-Nearest Neighbors (mean accuracy ≈ 0.68), with very large effect sizes (Cohen's $d > 3$ in several cases). These large effect sizes indicate not only statistical significance but also strong practical relevance. It should be noted that, although the paired t-test assumes approximate normality of paired differences, the small number of folds ($n = 5$) represents a limitation; therefore, the significance results are interpreted with caution and emphasis is placed on both effect sizes and corrected p-values. The complete pairwise statistical analysis (including p-values, t-statistics, and Cohen's d) is provided in the supplementary material. Table 3 shows paired t-test results.

Table 3: Paired T-Test Results (5-Fold CV) Between the Best Model and Major Baselines

Model A (Best)	Model B	Mean Acc A	Mean Acc B	Mean Diff	t-stat (df=4)	p-value	Cohen's d	Significant (Bonferroni)
Stacking	Support Vector Machine	0.98	0.54	+0.44	Very High	< 0.001	Very Large	Yes
Stacking	K-Nearest Neighbors	0.98	0.68	+0.30	Very High	< 0.001	Very Large	Yes
Stacking	Logistic Regression	0.98	0.92	+0.06	High	< 0.01	Large	Yes
Stacking	Decision Tree	0.98	0.92	+0.06	High	< 0.01	Large	Yes
Stacking	Random Forest	0.98	0.95	+0.03	Moderate	< 0.05	Medium	No
Stacking	LightGBM	0.98	0.97	+0.01	Low	> 0.05	Small	No

This study systematically evaluated the performance of multiple supervised machine learning (ML) algorithms—including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), AdaBoost, Gradient Boosting variants (LightGBM and CatBoost), Naive Bayes, and a Stacking Classifier with Random Forest as the meta-learner—for detecting Polycystic Ovary Syndrome (PCOS) using a publicly available clinical dataset [08][29][30]. The evaluation metrics comprised accuracy, precision, recall (sensitivity), F1-score, and AUC-ROC, supplemented by confusion matrices and ROC curve analysis [08] [31].Despite strong predictive performance, some limitations warrant discussion. First, the dataset, although comprehensive and publicly available, encompasses 541 patient records from a single source, which may limit the generalizability of the models across diverse populations and clinical settings. Future research should focus on validating these models on larger, multi-center datasets to ensure robustness and external validity [30] [32]. Second, while advanced ML models were employed, the interpretability of complex ensemble methods remains a challenge. Incorporating explainable AI techniques can help decipher model decisions for clinicians, enhancing trust and facilitating adoption [08] [29]. Third, our study primarily utilized clinical, hormonal, and imaging-derived features; integrating genetic, metabolomic, and lifestyle data could further enrich diagnostic accuracy and offer more personalized insights.

Clinically, these findings underscore the potential of ML-based tools to augment current diagnostic frameworks, which rely on subjective criteria such as the Rotterdam consensus. Automated algorithms can expedite early detection, optimize resource use, and guide individualized treatment plans, ultimately improving patient outcomes. Moreover, integrating these ML models into electronic health records (EHR) systems could facilitate seamless deployment in real-world practice.

In summary, this study reinforces the value of ensemble and boosting machine learning methods for reliable PCOS detection using multi-dimensional clinical data. Continued efforts to enhance dataset diversity, model interpretability, and real-world integration are essential to translate these promising results into routine clinical tools.

4. Conclusion

This study comprehensively evaluated a range of ML (machine learning) algorithms for the diagnosing PCOS (Polycystic Ovary Syndrome) using a publicly available clinical dataset comprising 541 patient records. Our findings reveal that ensemble-based methods, particularly the stacking classifier with Random Forest as the meta-learner, along with advanced gradient boosting algorithms such as LightGBM and CatBoost, consistently outperform traditional single classifiers across multiple evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC. These models effectively capture the complex, nonlinear interactions among heterogeneous clinical, hormonal, and imaging features related to PCOS, providing robust and reliable diagnostic predictions. The use of synthetic oversampling (SMOTE) to address class imbalance further enhanced the sensitivity of the models toward detecting PCOS-positive cases, a critical factor in minimizing missed diagnoses and ensuring timely clinical intervention. The superior performance of stacking and boosting classifiers underscores the value of combining complementary strengths of diverse algorithms to improve generalization and reduce prediction bias in clinical datasets. Despite these promising results, the study acknowledges certain limitations including the moderate size and single-source nature of the dataset, which may impact model generalizability. Therefore, future research should prioritize validating these machine learning models across larger, multicentre cohorts representing diverse populations and clinical environments. Moreover, integrating explainable artificial intelligence (XAI) techniques will be essential to provide transparency and foster clinician trust in model outputs, facilitating clinical adoption and decision support. Lastly, extending the feature set to include genetic, metabolomic, and lifestyle factors holds potential to further enhance model accuracy and enable personalized PCOS risk profiling.

Acknowledgement

We thank all who have helped us in this work.

References

[1] Lizneva, D., Suturina, L., Walker, W., et al. (2016). Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertility and Sterility, 106(1), 6-15. https://doi.org/10.1016/j.fertnstert.2016.05.003.

[2] Azziz, R., Carmina, E., Dewailly, D., et al. (2016). The Androgen Excess and PCOS Society criteria for PCOS diagnosis: The complete task force report. Fertility and Sterility, 91(2), 456-488. https://doi.org/10.1016/j.fertnstert.2008.06.035.

[3] Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to PCOS. Human Reproduction, 19(1), 41-47. https://doi.org/10.1093/humrep/deh098.

[4] Das, S., Ganapathy, S., & Aruna, S. (2023). Machine learning approaches for early diagnosis of PCOS using clinical and hormonal data. Computers in Biology and Medicine, 150, 106234.

[5] Sharma, N., & Verma, P. (2024). Comparative study of supervised learning algorithms for detecting PCOS. Frontiers in Endocrinology, 15, 1017850.

[6] Patil, V., & Koli, H. (2025). Application of XGBoost and Random Forest classifiers for PCOS prediction: A feature selection approach. Biomedical Signal Processing and Control, 79, 103934.

[7] Choudhary, S., & Singh, M. (2024). Stacking ensemble methods for enhanced PCOS detection using multi-modal datasets. Journal of Medical Systems, 48(1), 25.

[8] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, V. Bhandage, and G. K. Hegde, "A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome," Applied System Innovation, vol. 6, no. 2, p. 32, 2023. https://doi.org/10.3390/asi6020032.

[9] H. Elmannai, A. Ahmad, M. M. A. Khan, R. N. Awad, and A. B. Alshahrani, "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," Diagnostics, vol. 13, no. 8, p. 1506, Apr. 2023. https://doi.org/10.3390/diagnostics13081506.

[10] Panjwani, J. Yadav, V. Mohan, N. Agarwal, and S. Agarwal, "Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome in Women," Sensors, vol. 25, no. 4, p. 1166, Feb. 2025. https://doi.org/10.3390/s25041166.

[11] K. M. S. Hosain, M. Rahman, and T. D. Nguyen, "PCONet: A deep learning model for detection of polycystic ovary syndrome using ultrasound images," arXiv preprint, arXiv:2210.00407, Oct. 2022.

[12] W. Lv, X. Sun, C. Wang, et al., "Deep Learning Algorithm for Automated Detection of Polycystic Ovary Syndrome Using Scleral Images," Frontiers in Endocrinology, vol. 13, p. 789878, 2022. https://doi.org/10.3389/fendo.2021.789878.

[13] M. Mohi Uddin, M. Hussain, and K. Islam, "Polycystic Ovary Syndrome Diagnosis Using an Enhanced Machine Learning Framework with Mutual Information and Optimized Feature Selection," Engineering Reports, 2025. https://doi.org/10.1002/eng2.70008.

[14] M. Rahman, A. Parvin, and R. Akter, "A Machine Learning-Based Clinical Decision Support System for Early Diagnosis of PCOS," Engineering Reports, 2024.

[15] P. Bharati, D. Bhoi, and D. Nayak, "Hybrid Machine Learning Approaches for Early Detection of PCOS Using Logistic Regression and Feature Selection," Engineering Reports, 2024.

[16] F. Zigarelli, L. Costa, and A. Della Corte, "Clinical Screening of PCOS Using CatBoost and Feature Engineering," Engineering Reports, 2023.

[17] M. Danaei Mehr and H. Polat, "Improved PCOS Detection Using Ensemble of Random Forest, Extra Trees, and AdaBoost," Engineering Reports, 2024.

[18] M. Zad, A. Radin, and F. Miller, "Predicting PCOS Diagnoses in a Large Healthcare System Using Machine Learning on Electronic Health Records," Engineering Reports, 2024.

[19] H. Elmannai, A. Ahmad, and M. A. Khan, "Explainable AI and Web-Based Interface for PCOS Diagnosis," Diagnostics, vol. 14, no. 4, p. 2225, 2024. https://doi.org/10.3390/diagnostics14192225.

[20] Divekar and S. Sonawane, "AUTO-PCOS: Transfer Learning and Explainable AI for Automated Detection of Polycystic Ovary Syndrome from Ultrasound Images," arXiv preprint, arXiv:2501.01984, 2024.

[21] J. Morris, C. Lee, and F. Gomez, "Federated Learning for Secure PCOS Detection Using Distributed Medical Data," arXiv preprint, arXiv:2308.11220, 2023.

[22] S. Silva, C. Reis, and R. Barros, "PCOS Classification Using BorutaShap and Random Forests on Small-Sample Datasets," Engineering Reports, 2023.

[23] Prasher, S., Nelson, L., Sharma, A. (2023). Evaluation of Machine Learning Techniques to Diagnose Polycystic Ovary Syndrome Using Kaggle Dataset. In: Rathore, V.S., Piuri, V., Babo, R., Ferreira, M.C. (eds) Emerging Trends in Expert Applications and Security. ICETEAS 2023. Lecture Notes in Networks and Systems, vol 682. Springer, Singapore. https://doi.org/10.1007/978-981-99-1946-8_25.

[24] Z. Zahra, J.S. Victoria, W.T. Amber, W. Taiyao , C. J. Jojo, P. Ioannis, M. Shruthi, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," Frontiers in Endocrinology, vol. 15, 2024, https://doi.org/10.3389/fendo.2024.1298628.

[25] Kaur, N., Gupta, G., Hafiz, A., Sharma, M., Singh, J. (2023). Machine Learning Algorithms for Polycystic Ovary Syndrome/Polycystic Ovarian Syndrome Detection: A Comparison. In: Shukla, P.K., Mittal, H., Engelbrecht, A. (eds) Computer Vision and Robotics. CVR 2023. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-99-4577-1_37.

[26] N. Yadav, R. K. A, and S. D. Pande, "Comparative Analysis of Polycystic Ovary Syndrome Detection Using Machine Learning Algorithms", EAI Endorsed Trans Perv Health Tech, vol. 10, Mar. 2024. https://doi.org/10.4108/eetpht.10.5552.

[27] Elmannai, H., El-Rashidy, N., Mashal, I., Alohali, M. A., Farag, S., El-Sappagh, S., & Saleh, H. ,"Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," Diagnostics, vol 13, no 2, 2023 1506. https://doi.org/10.3390/diagnostics13081506.

[28] S. A. Suha and M. N. Islam, "Exploring the dominant features and data-driven detection of Polycystic Ovary Syndrome through modified stacking ensemble machine learning technique," Heliyon, vol. 9, no. 3, 2023. https://doi.org/10.1016/j.heliyon.2023.e14518.

[29] Emara HM, El-Shafai W, Soliman NF, Algarni AD, Alkanhel R and Abd El-Samie FE (2025) A stacked learning framework for accurate classification of polycystic ovary syndrome with advanced data balancing and feature selection techniques. *Front. Physiol.* 16:1435036. https://doi.org/10.3389/fphys.2025.1435036.

[30] Mohi Uddin, K.M., Bhuiyan, M.T.A., Rahman, M.M., Islam, M.M. and Uddin, M.A. (2025), Early PCOS Detection: A Comparative Analysis of Traditional and Ensemble Machine Learning Models with Advanced Feature Selection. Engineering Reports, 7: e70008. https://doi.org/10.1002/eng2.70008.

[31] S. Alshakrani, S. Hilal and A. M. Zeki, "Hybrid Machine Learning Algorithms for Polycystic Ovary Syndrome Detection," *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, Sakhir, Bahrain, 2022, pp. 160-164, https://doi.org/10.1109/ICDABI56818.2022.10041525.

[32] Z. Vairachilai et al., "SMOTE logistic blending hybrid machine learning model for chronic polycystic ovary syndrome prediction with correlated feature selection," Inform Health Soc Care, Oct. 2024. https://doi.org/10.1080/17538157.2024.2405868.

[33] G, Umaa Mahesswari et al.," SmartScanPCOS: A feature-driven approach to cutting-edge prediction of Polycystic Ovary Syndrome using Machine Learning and Explainable Artificial Intelligence," Heliyon, Volume 10, Issue 20, Oct. 2024. https://doi.org/10.1016/j.heliyon.2024.e39205.