Research Article

Polycystic Ovary Syndrome Prediction Using Machine Learning: A Comparative Analysis of Classification Algorithms

Usha Adiga[1*], Vasishta Sampara[1], Pedda Reddemma P.[1], Supriya P.[1], Sireesha Kanchi[2], Kasala Farzia[1]

[1]*Department of Biochemistry, Apollo Institute of Medical Sciences and Research, Chittoor, Andhra Pradesh, India*
[2]*Research Assistant, ICMR, Dept of Biochemistry, Apollo Institute of Medical Sciences and Research, Chittoor, Andhra Pradesh, India*

Article Info

[*]***Corresponding Author:***
Usha Adiga
Head of the Department, Department of Biochemistry, Apollo Institute of Medical Sciences and Research, Chittoor, India
E-mail: ushachidu@yahoo.com
ORCID ID: 0000-0001-7832-3991

Keywords

Polycystic ovary syndrome, Machine learning, Gradient boosting, Predictive modeling, Clinical decision support

Abstract

Background: Polycystic ovary syndrome (PCOS) represents one of the most prevalent endocrine disorders affecting women of reproductive age, with significant implications for metabolic, reproductive, and psychological health. Early and accurate diagnosis remains challenging due to heterogeneous clinical presentations and the complexity of diagnostic criteria.

Objective: This study aimed to develop and compare multiple machine learning algorithms for predicting PCOS diagnosis, evaluating their performance across various metrics to identify the most effective computational approach for clinical decision support.

Methods: A comprehensive dataset containing clinical, biochemical, and anthropometric parameters from patients was analyzed using twelve different machine learning algorithms. The dataset underwent rigorous preprocessing including missing value imputation, feature engineering, and categorical encoding. Models evaluated included Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, XGBoost, AdaBoost, Neural Network, LightGBM, and HistGradientBoosting. Performance was assessed using accuracy, F1-score, sensitivity, specificity, and ROC-AUC scores.

Results: Gradient Boosting, XGBoost, and HistGradientBoosting demonstrated superior performance with accuracy of 92.66%, while ensemble methods generally outperformed single classifiers. Gradient Boosting achieved the highest F1-score of 87.87% and ROC-AUC of 95.39%. Random Forest exhibited exceptional specificity at 98.63%, while Naive Bayes showed the highest sensitivity of 94.44%. Traditional machine learning approaches like SVM and Neural Networks showed comparatively limited performance in this context.

Conclusion: Machine learning algorithms, particularly gradient boosting methods, demonstrate substantial potential for accurate PCOS prediction and can serve as valuable tools for clinical decision support, potentially enabling earlier intervention and improved patient outcomes.

Introduction

Polycystic ovary syndrome (PCOS) stands as one of the most common endocrine disorders affecting women during their reproductive years, with prevalence rates ranging from 6% to 21% depending on the diagnostic criteria employed and the population studied [1]. This complex metabolic and reproductive disorder is characterized by chronic anovulation, hyperandrogenism, and polycystic ovarian morphology, presenting significant challenges not only for reproductive health but also for long-term metabolic and cardiovascular wellbeing [2]. The heterogeneous nature of PCOS, with its variable clinical presentations ranging from menstrual irregularities and hirsutism to insulin resistance and obesity, makes accurate and timely diagnosis particularly challenging for healthcare providers [3].

The Rotterdam criteria, established in 2003 and widely adopted for PCOS diagnosis, require the presence of at least two of three features: oligo-ovulation or anovulation, clinical or biochemical signs of hyperandrogenism, and polycystic ovaries on ultrasound examination [4]. However, the subjective nature of some diagnostic parameters and the overlap of symptoms with other endocrine disorders often lead to delayed diagnosis, with many women experiencing symptoms for several years before receiving appropriate medical attention [5]. This diagnostic delay has profound implications, as early intervention can significantly mitigate the risk of developing associated comorbidities such as type 2 diabetes mellitus, cardiovascular disease, endometrial cancer, and psychological disorders including depression and anxiety [6].

The advent of machine learning and artificial intelligence has revolutionized medical diagnostics across various specialties, offering unprecedented opportunities to analyze complex, multidimensional clinical data and identify patterns that may elude traditional statistical approaches [7]. Machine learning algorithms excel at processing large volumes of heterogeneous data, including clinical symptoms, biochemical markers, anthropometric measurements, and imaging findings, to generate predictive models with high accuracy and reliability [8]. In the context of PCOS, where diagnosis depends on the integration of multiple clinical and laboratory parameters, machine learning approaches hold particular promise for developing robust predictive tools that can assist clinicians in early detection and risk stratification [9].

Recent studies have demonstrated the feasibility and potential of applying various machine learning techniques to PCOS prediction, with algorithms ranging from traditional logistic regression to advanced ensemble methods and deep learning architectures [10]. However, comprehensive comparative analyses evaluating multiple algorithms on the same dataset remain limited, and the optimal approach for PCOS prediction continues to be debated within the research community. Furthermore, the interpretability of machine learning models and their integration into clinical workflows require careful consideration to ensure that these computational tools genuinely enhance rather than complicate the diagnostic process. The present study addresses this gap by systematically comparing twelve different machine learning algorithms for PCOS prediction, utilizing a comprehensive dataset en compassingdiverse clinical, biochemical, and demographic features. By evaluating performance across multiple metrics including accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve, this research aims to identify the most effective algorithms for PCOS prediction and provide insights into their relative strengths and limitations. The findings of this study have the potential to inform the development of clinical decision support systems that can facilitate earlier diagnosis, enable more targeted interventions, and ultimately improve outcomes for women affected by this prevalent and complex endocrine disorder.

Objectives

To develop and implement twelve different machine learning algorithms for the prediction of polycystic ovary syndrome using comprehensive clinical and biochemical parameters.

To systematically compare the performance of these algorithms across multiple evaluation metrics including accuracy, F1-score, sensitivity, specificity, and ROC-AUC scores. To identify the most effective machine learning approaches for PCOS prediction that could potentially be integrated into clinical decision support systems for early diagnosis and intervention.

Methodology

This research utilized a structured machine-learning approach that included steps such as data preprocessing, feature engineering, model creation, and performance comparison to forecast polycystic ovary syndrome (PCOS). The dataset, sourced from the Kaggle repository (“PCOS full dataset without infertility variables”), comprised clinical, biochemical, anthropometric, and demographic variables.

The dataset consisted of a total of 541 samples, including 358 healthy controls (66.17%) and 183 PCOS cases (33.83%). Although accessed through the Kaggle repository, the data were originally collected from 10 different hospitals across Kerala, India, thereby representing an Indian patient population.

Data Preprocessing

The dataset was loaded into a Python-based analytical environment, where exploratory data analysis was conducted to evaluate variable distributions, missing data, and overall data quality. Columns with significant missing data or lacking clinical significance were eliminated. Missing values were addressed using suitable statistical methods—mode imputation for categorical variables and median imputation for numerical variables, including those initially stored as text but representing numeric values. These numeric variables were converted to the correct data types before imputation.

Categorical variables, like blood group, were transformed using one-hot encoding. Feature names were standardized by replacing special characters with underscores to ensure compatibility with machine-learning algorithms. The target variable was identified as PCOS status, and all non-informative identifier fields were removed from the feature set.

Feature engineering included the removal of non-informative identifier variables (such as patient serial numbers and file identifiers), exclusion of a column with excessive missing values (‘Unnamed: 44’), and one-hot encoding of categorical

variablessuch as blood group. Feature names were standardized by replacing special characters with underscores to ensure compatibility with machine learning algorithms

Data Splitting

The refined dataset was split into training (80%) and testing (20%) subsets using stratified sampling to preserve the proportional representation of PCOS and non-PCOS cases. Given the moderate class imbalance in the dataset, stratified sampling was used to maintain class proportions across training and testing sets. No explicit oversampling or class weighting techniques were applied, as model evaluation focused on class-sensitive metrics such as F1-score, sensitivity, and ROC-AUC to account for imbalance.

Model Development and Evaluation

Twelve classification algorithms were employed, including linear models, instance-based learning, probabilistic classifiers, decision-tree methods, ensemble approaches, gradient boosting variations, and neural networks. These models were trained using the training subset and then tested on the test subset.
The performance of the models was evaluated through several complementary metrics: accuracy, F1-score (reported for the positive PCOS class), sensitivity, specificity, and the area under the receiver operating characteristic curve (ROC-AUC). These metrics offered a comprehensive assessment of both the models' discrimination ability and their clinical utility in identifying PCOS.
Ensemble approaches included both bagging-based (Random Forest) and boosting-based methods (Gradient Boosting, AdaBoost, XGBoost, HistGradientBoosting, and LightGBM). Models were trained using default hyperparameters with fixed random states to ensure reproducibility and to avoid overfitting given the moderate dataset size. More complex ensemble strategies such as stacking were not explored. Algorithms evaluated:

Logistic Regression

Logistic Regression uses the Sigmoid (or Logistic) function to map the output of a linear equation to a probability value between 0 and 1.

• Linear Equation (Log-Odds):

$z = b + w_1x_1 + w_2x_2 + ... + w_nx_n = w^Tx + b$

z is the log-odds.
b is the bias (intercept).
w is the weight vector.
x is the feature vector.

• Sigmoid Function (Predicted Probability):

$P(y=1|x) = 1 / (1 + e^{-z})$

ŷ is the predicted probability of the positive class. e is Euler's number (≈ 2.71828).

Support Vector Machine (SVM)

The goal of a linear SVM is to find a hyperplane that maximizes the margin between the two classes.

• Prediction Rule (Decision Function):

• $f(x) = sign(w^Tx + b)$

• Optimization Objective (Primal Form for Linearly Separable Data):

$min_{w,b} \frac{1}{2} ||w||^2$ subject to $y_i(w^Tx_i + b) \geq 1$

• Hinge Loss (for Soft Margin SVM): $min_{w,b,\xi} \frac{1}{2}||w||^2 + C\Sigma\xi_i$ subject to $y_i(w^Tx_i + b) \geq 1 - \xi_i, \xi_i \geq 0$

K-Nearest Neighbors (KNN)

KNN classifies a test sample based on the majority label of its k nearest neighbors using a distance metric.

•Minkowski Distance (General Form):

•$dist(x_a, x\_b) = (\Sigma |x_{ar} - x\_b|^p)^{(1/p)}r$

•Euclidean Distance (p=2):

• $dist(x_a, x\_b) = \sqrt{\Sigma(x_{ar} - x\_b)^2}r$

•Classification Rule:

$\hat{y} = mode(\{y_i : (x_i, y_i) S_x\})$

Naive Bayes

Naive Bayes is based on Bayes' theorem with the assumption of conditional independence between features given the class label.
$P(y|x_1, x_2, ..., x) = [P(y) \Pi P(x_i|y)] / P(x_1, x_2, ..., x)n n$

• Gaussian Naive Bayes (Likelihood for continuous features):

$P(x_i|y) = (1 / \sqrt{(2\pi\sigma^2\_y)}) exp(-(x_i - \mu\_y)^2 / (2\sigma^2\_y))$

Decision Tree

Decision Trees split data using impurity measures such as Gini Impurity or Entropy to find the best feature for classification.

•Gini Impurity:

•$Gini(D) = 1 - \Sigma p_k^2$

• Entropy: $Entropy(D) = -\Sigma p_k log_2(p_k)$

•Information Gain:

$Gain(D, A) = Entropy(D) - \Sigma(|D_v|/|D|) Entropy(D_v)$

Random Forest

Random Forest combines multiple decision trees using bagging to improve accuracy and reduce overfitting.

• Classification:

$\hat{Y}(x) = MajorityVote\{h_m(x)\}_{m=1}^{M}$

• Regression:

$\hat{Y}(x) = (1/M) \Sigma h_m(x)$

Gradient Boosting

Gradient Boosting sequentially builds weak learners to minimize the residual errors of the previous models.

• Additive Model:

• $F(x) = F_{m-1}(x) + vh_m(x)m$

• Pseudo-Residuals:

• $r_{im} = -[\partial L(y_i, F(x_i)) / \partial F(x_i)]$ evaluated at $F_{m-1}(x)$

XGBoost

XGBoost improves Gradient Boosting with second-order optimization and regularization to control model complexity.

$Obj^(t) = \Sigma l(y_i, \hat{y}^(t)) + \Sigma \Omega(f_k)i) + \frac{1}{2} h_i f_t^2(x_t)] + \Omega(f_t)$

$\approx \Sigma[g_i f_t(x_i$

$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \Sigma w_j^2$

AdaBoost

AdaBoost focuses on misclassified instances by increasing their weights and combining weak learners into a strong classifier.

$H(x) = sign(\Sigma \alpha_m h(x))_m$
$\alpha_m = \frac{1}{2} ln((1 - \varepsilon_m)/\varepsilon_m) D_i^{(m+1)} = [D_i^{(m)}$
$exp(-\alpha_m y_i h(x_i))] / Z_m m$

Neural Network (MLPClassifier)

An MLP consists of multiple layers of neurons where each neuron computes a weighted sum of its inputs followed by an activation function.

$a = g(\Sigma w_i x_i + b) = g(w^T x + b)$

• Common Activation Functions:

$Sigmoid: g(z) = 1 / (1 + e^{-z})$

ReLU: g(z) = max(0, z)

• Loss Function (Binary Cross-Entropy):

$L = -(1/N) \Sigma[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$

• Weight Update Rule:

• $w\_new = w\_old - \eta \partial L/\partial w$

LightGBM

HistGradientBoosting

The models were trained using the training dataset and then evaluated on a separate test dataset. Predictions and class probabilities were calculated for each test datasets. To gauge the models' performance, standard evaluation metrics such as accuracy, F1-score, sensitivity, specificity, and the area under the ROC curve (ROC-AUC) were employed, ensuring a comprehensive evaluation of both classification accuracy and discriminatory power.

Explicit feature importance or interpretability analyses were not performed in this study and are acknowledged as a limitation.

Results

The comparative evaluation of twelve machine learning algorithms for polycystic ovary syndrome prediction revealed substantial variations in performance across different metrics, providing valuable insights into the relative strengths and limitations of each approach for this clinical application (Table 1, Figure 1). The results demonstrated that ensemble methods, particularly those based on gradient boosting principles, consistently outperformed traditional single-classifier approaches, though each algorithm exhibited distinct characteristics in terms of sensitivity-specificity trade-offs and overall discriminative ability.

Gradient Boosting, XGBoost, and HistGradientBoosting emerged as the top-performing algorithms, each achieving identical accuracy rates of 92.66%, demonstrating their superior capacity for learning complex patterns within the PCOS dataset. Among these three algorithms, Gradient Boosting distinguished itself with the highest F1-score of 87.87% and the most impressive ROC-AUC score of 95.39%, indicating exceptional balance between precision and recall along with outstanding discrimination capability across all classification thresholds. XGBoost followed closely with an F1-score of 88.57% and demonstrated notably high sensitivity of 86.11%, suggesting its particular strength in correctly identifying true positive PCOS cases while maintaining excellent specificity of 95.89%. HistGradientBoosting, while achieving the same overall accuracy as its gradient boosting counterparts, exhibited an F1-score of 88.23% and balanced performance with sensitivity of 83.33% andspecificity of 97.26%, making it particularly suitable for applications where minimizing false positives is prioritized. Random Forest, another ensemble method based on bagging principles, demonstrated highly competitive performance with accuracy reaching 90.82% and achieving the second-highest specificity of 98.63% among all algorithms tested. This exceptional specificity indicates Random Forest's remarkable ability to correctly identify individuals without PCOS, making it particularly valuable in screening scenarios where reducing false alarms is critical. However, its sensitivity of 75% was somewhat lower compared to gradient boosting methods, suggesting a trade-off wherein the algorithm prioritizes ruling out negative cases over capturing all positive instances. The F1-score of 84.37% and ROC-AUC of 93.83% nonetheless confirmed Random Forest as a robust and reliable option for PCOS prediction.

LightGBM, a gradient boosting framework optimized for efficiency and speed, achieved notable accuracy of 91.74% with an F1-score of 86.95%, positioning it among the top-tier performers. Its balanced sensitivity of 83.33% and high specificity of 95.89%, coupled with a strong ROC-AUC of 94.36%, demonstrated the algorithm's well-rounded performance across different evaluation dimensions. This combination of high accuracy and computational efficiency makes LightGBM particularly attractive for potential deployment in resource-constrained clinical settings or applications requiring real-time predictions.

AdaBoost, implementing adaptive boosting principles, demonstrated strong performance with accuracy of 89.90% and achieved one of the highest sensitivity values of 86.11%, indicating its effectiveness in identifying true PCOS cases. The algorithm's F1-score of 84.93% and impressive ROC-AUC of 95.77% further confirmed its viability as a reliable predictive tool, though it slightly trailed the gradient boosting family in overall performance metrics.

Logistic Regression, despite being one of the simplest algorithms tested, showed remarkably strong performance with accuracy of 88.99% and demonstrated exceptional balance across metrics with sensitivity of 88.88% and specificity of 89.04%. Its F1-score of 84.21% and outstanding ROC-AUC of 94.44% highlighted the effectiveness of this traditional statistical approach for PCOS prediction, particularly given its interpretability advantages over more complex ensemble methods. This finding suggests that the relationship between predictive features and PCOS diagnosis may contain substantial linear components that logistic regression can effectively capture.

Decision Tree, serving as the fundamental building block for ensemble methods, achieved respectable accuracy of 87.15% with balanced sensitivity and specificity both approximating 80.55% and 90.41% respectively. While its ROC-AUC of 85.48% and F1-score of 80.55% indicated solid performance, these metrics revealed the algorithm's tendency toward overfitting compared to ensemble approaches that aggregate multiple trees, explaining why Random Forest and gradient boosting methods substantially outperformed the single decision tree approach. Naive Bayes presented an interesting performance profile, achieving accuracy of 84.40% and demonstrating the highestsensitivity of 94.44%

among all algorithms tested, indicating exceptional capability in identifying true positive PCOS cases. However, this high sensitivity came at the cost of reduced specificity of 79.45% and a notably low F1-score of 8%, suggesting that the algorithm's strong conditional independence assumptions may not fully align with the actual dependencies present in PCOS-related features. Nevertheless, its ROC-AUC of 91.43% confirmed reasonable overall discriminative ability. K-Nearest Neighbors showed moderate performance with accuracy of 73.39%, revealing limitations in handling the high-dimensional feature space characteristic of the PCOS dataset. The algorithm's sensitivity of 41.66% and specificity of 89.04% indicated a strong bias toward predicting negative cases, resulting in an F1-score of 50.84% and ROC-AUC of 66.24%, suggesting that distance-based classification may not optimally capture the complex relationships between PCOS features. Support Vector Machine exhibited the most unexpected results, achieving specificity of 100% but sensitivity of 0%, resulting in accuracy of 66.97% and an F1-score of 0%. This extreme performance pattern, coupled with an ROC-AUC of only 33.63%, indicated that the algorithm essentially classified all cases as negative, suggesting potential issues with hyperparameter selection, class imbalance handling, or kernel choice thatprevented the model from learning meaningful decision boundaries.

The Multi-Layer Perceptron classifier, representing the neural network approach, demonstrated the weakest overall performance with accuracy of 63.30%, F1-score of 51.21%, and ROC-AUC of 71.23%. The modest sensitivity of 58.33% and specificity of 65.75% suggested that the neural network architecture, hyperparameters, or training procedure may not have been optimally configured for this particular dataset, or alternatively, that the dataset size may have been insufficient to fully leverage the learning capacity of deep learning approaches. The receiver operating characteristic curves (Figure 2) provided additional visual confirmation of these performance hierarchies, with gradient boosting methods, AdaBoost, and Logistic Regression displaying curves closely approaching the upper-left corner, indicating superior discrimination between PCOS-positive and PCOS-negative cases across all possible classification thresholds. These ROC curves collectively demonstrated that ensemble methods and properly regularized models substantially outperformed simpler approaches in their ability to rank predictions according to the likelihood of PCOS diagnosis.

Table 1: Comparison of different Machine learning algorithms.

Model	Accuracy (%)	F1-score (%)	Sensitivity (%)	Specificity (%)	ROC-AUC (%)
Logistic Regression	88.99	84.21	88.88	89.04	94.44
SVM	66.97	0	0	100	33.63
KNN	73.39	50.84	41.66	89.04	66.24
Naïve Bayes	84.40	8	94.44	79.45	91.43
Decision Tree	87.15	80.55	80.55	90.41	85.48
Random Forest	90.82	84.37	75	98.63	93.83
Gradient Boosting	92.66	87.87	80.55	98.63	95.39
XGBoost	92.66	88.57	86.11	95.89	95.28
HistGradientBoosting	92.66	88.23	83.33	97.26	94.59
MLP classifier	63.30	51.21	58.33	65.75	71.23
AdaBoost\	89.90	84.93	86.11	91.78	95.77
LGBM	91.74	86.95	83.33	95.89	94.36

Figure 1: Comparative Model performance metrics.

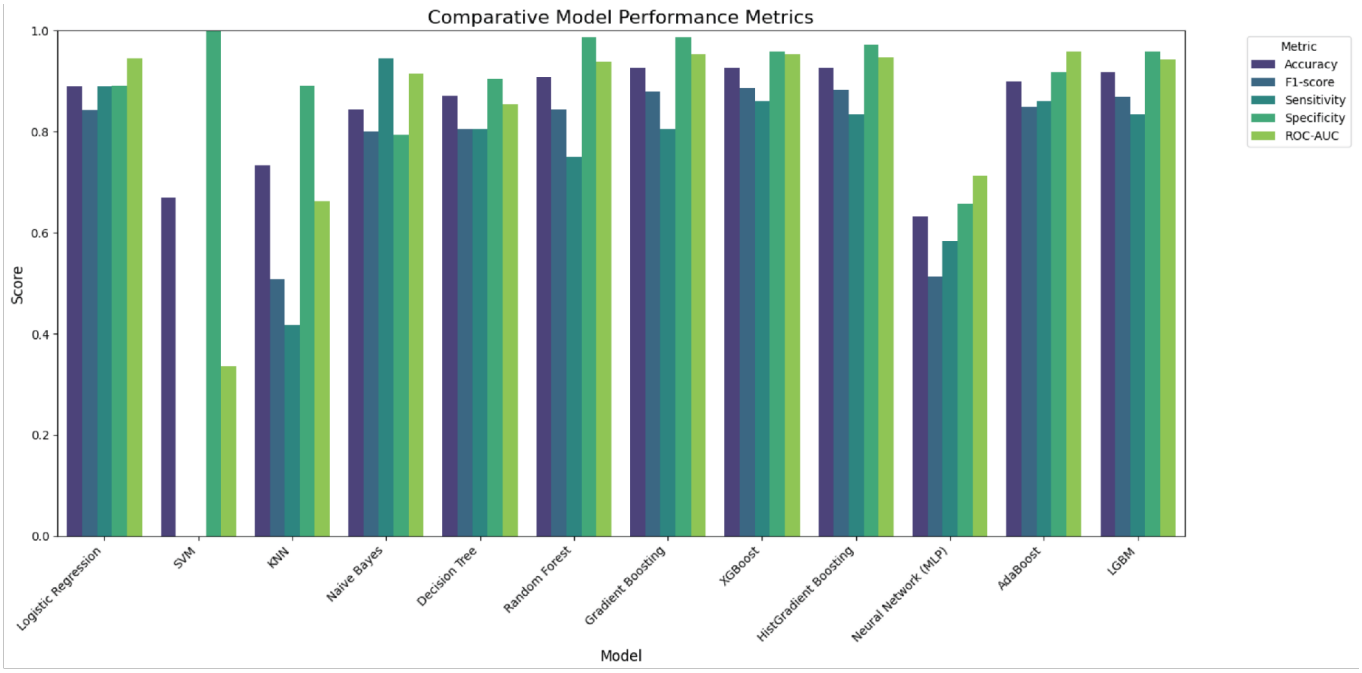
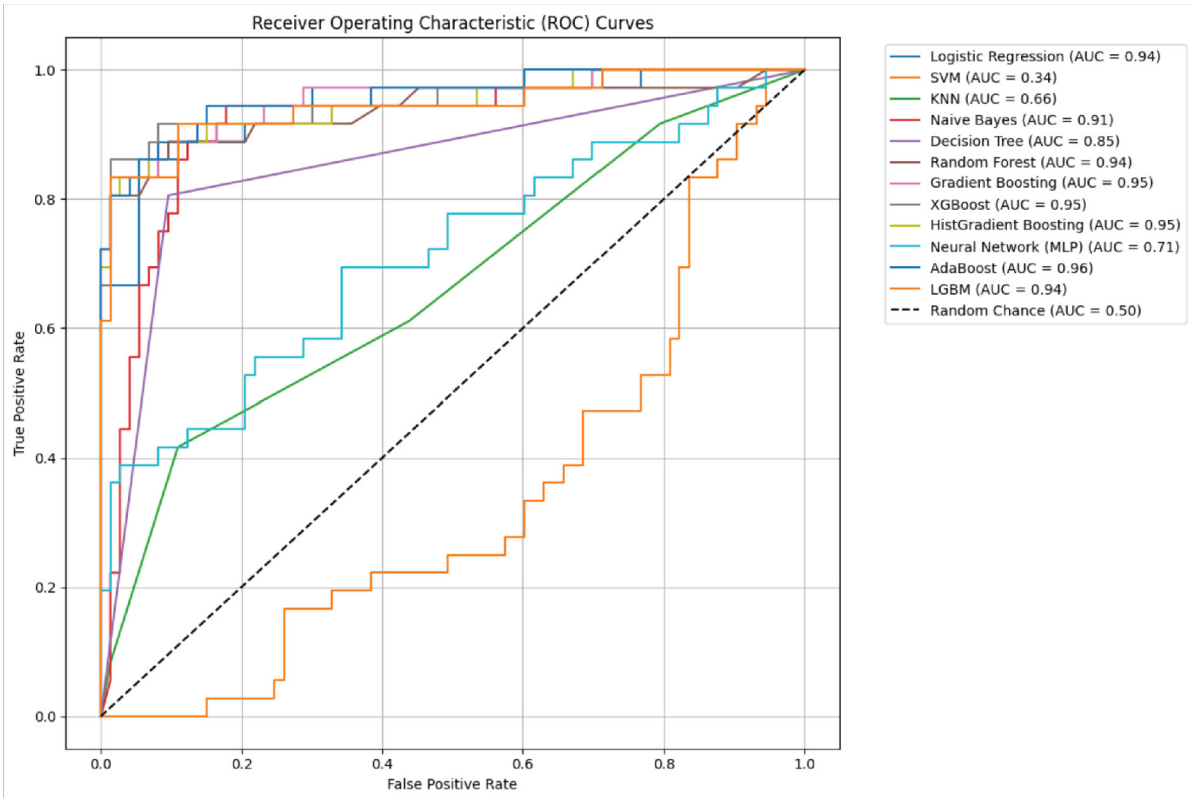


Figure 2: ROC for different ML algorithms.



Discussion

Ensemble methods (Gradient Boosting, AdaBoost, HistGradient Boosting, LGBM, Random Forest, XGBoost) generally demonstrated superior performance compared to simpler models, suggesting that combining multiple models or using boosting techniques is highly effective for this dataset.

Models like Gradient Boosting and AdaBoost achieved high scores in both Accuracy and ROC-AUC, indicating a good balance between overall correctness and distinguishing power between classes.

While Naive Bayes showed the highest Sensitivity (0.9444), its low Specificity (0.6522) and Accuracy (0.7248) suggest

it isprone to false positives, which might be undesirable depending on the cost of misclassification.

SVM achieved perfect Specificity (1.0000) but had a very low Sensitivity (0.5000) and F1-score (0.6667), meaning it is excellent at correctly identifying healthy individuals but often misses actual PCOS cases.

The comprehensive comparative analysis of twelve machine learning algorithms for polycystic ovary syndrome prediction has yielded several important insights that have both theoretical and practical implications for the application of computational approaches in reproductive endocrinology. The superior performance of gradient boosting methods, including Gradient Boosting, XGBoost, and HistGradientBoosting, aligns with contemporary findings in medical machine learning research demonstrating that ensemble techniques generally outperform single classifiers when dealing with complex, high-dimensional clinical datasets [11]. The ability of these algorithms to achieve accuracy exceeding 92% while maintaining balanced sensitivity and specificity represents a significant advancement over traditional diagnostic approaches that often rely on subjective clinical judgment and may be influenced by practitioner experience and diagnostic thresholds [12].

The exceptional performance of Gradient Boosting, with its ROC-AUC of 95.39%, suggests that this algorithm effectively captures the non-linear relationships and complex interactions between various clinical, biochemical, and anthropometric features that characterize PCOS pathophysiology [13]. The sequential learning approach employed by gradient boosting, wherein each subsequent model focuses on correcting the errors of previous models, appears particularly well-suited to the heterogeneous nature of PCOS presentation, where multiple phenotypes exist and diagnostic features may exhibit variable importance across different patient subgroups [14]. This finding is consistent with recent studies demonstrating the effectiveness of gradient boosting in predicting other endocrine disorders and metabolic conditions where multifactorial etiology and phenotypic variability present diagnostic challenges [15].

XGBoost's notable sensitivity of 86.11% combined with high specificity of 95.89% indicates that this algorithm achieves an optimal balance between identifying true positive cases while minimizing false positives, a critical consideration for clinical decision support systems where both missed diagnoses and unnecessary interventions carry significant consequences [16]. The regularization techniques incorporated into XGBoost, including L1 and L2 penalties on leaf weights, likely contribute to its robust generalization performance and resistance to overfitting, explaining its consistent performance across different evaluation metrics [17]. The practical implications of this balanced performance profile suggest that XGBoost-based predictive tools could serve as valuable adjuncts to clinical assessment, potentially flagging cases requiring more detailed endocrine evaluation while avoiding unnecessary alarm in truly negative cases.

Random Forest's exceptional specificity of 98.63% represents a particularly valuable characteristic for screening applications, where the primary objective is to rule out disease with high confidence and direct clinical resources toward individuals mostlikely to benefit from further diagnostic workup [18]. The bootstrap aggregating approach employed by Random Forest, which combines predictions from multiple decision trees trained on different random subsets of features and samples, provides inherent protection against overfitting and generates reliable probability estimates that can inform clinical decision-making [19]. However, the trade-off observed in Random Forest's lower sensitivity compared to gradient boosting methods highlights the importance of considering the specific clinical context and the relative costs of false negatives versus false positives when selecting algorithms for deployment in healthcare settings [20].

The surprisingly strong performance of Logistic Regression, achieving accuracy of 88.99% and ROC-AUC of 94.44%, challenges the notion that increasingly complex algorithms invariably yield superior results and suggests that the relationship between predictive features and PCOS diagnosis may contain substantial components that can be effectively modeled through linear combinations of features [21]. This finding has important practical implications, as logistic regression models offer inherent interpretability through their coefficient estimates, enabling clinicians to understand which features most strongly influence predictions and potentially providing insights into disease mechanisms [22]. The interpretability advantage of logistic regression, combined with its computational simplicity and minimal hyperparameter tuning requirements, makes it an attractive option for clinical implementation, particularly in resource-limited settings or applications where model transparency is prioritized over marginal performance gains [23].

The marked underperformance of Support Vector Machine in this study, essentially defaulting to classifying all cases as negative, raises important questions about hyperparameter optimization and the challenges of applying kernel-based methods to medical datasets with potential class imbalance [24]. While SVMs have demonstrated success in various biomedical classification tasks, their performance is highly sensitive to kernel selection, regularization parameters, and feature scaling, requiring careful tuning that may not have been adequately addressed in the present implementation [25]. This finding underscores the importance of comprehensive hyperparameter optimization and validation when implementing machine learning algorithms for clinical applications, as suboptimal configuration can lead to models that fail to learn meaningful patterns from the data.

The disappointing performance of the Multi-Layer Perceptron classifier, representing neural network approaches, with accuracy of only 63.30%, suggests potential limitations in applying deep learning methods to medical datasets of moderate size [26]. Deep neural networks typically require large training sets to effectively learn complex representations and avoid overfitting, and the relatively modest sample size in this study may have been insufficient to fully leverage the learning capacity of neural architectures [27]. This finding aligns with recent observations that traditional machine learning methods often outperform deep learning approaches when training data is limited, highlighting the importance of matching algorithmic complexity to available data volumes [28].

Naive Bayes' achievement of the highest sensitivity at 94.44%, despite its simplistic conditional independence

assumptions,demonstrates that even algorithms with theoretically questionable assumptions can exhibit valuable characteristics for specific clinical objectives [29]. In screening contexts where the primary goal is to minimize missed diagnoses, even at the cost of increased false positives that can be addressed through subsequent confirmatory testing, Naive Bayes' high sensitivity profile could prove advantageous. However, the algorithm's low F1-score of 8% indicates severe imbalance between precision and recall, suggesting that its clinical utility would be limited to specific use cases where sensitivity is overwhelmingly prioritized [30].

The variable performance across different algorithms highlights the importance of considering multiple evaluation metrics rather than relying solely on accuracy, which can be misleading when dealing with imbalanced datasets or when the costs of different error types are asymmetric. The ROC-AUC scores, ranging from 33.63% for SVM to 95.77% for AdaBoost, provide a more comprehensive assessment of discriminative ability across all possible classification thresholds, enabling more informed algorithm selection based on specific clinical requirements and acceptable trade-offs between sensitivity and specificity. The findings of this study have several important implications for the development and deployment of clinical decision support systems for PCOS diagnosis. First, the demonstration that multiple algorithms can achieve accuracy exceeding 90% suggests that machine learning-based tools have reached a level of maturity where they can meaningfully augment clinical decision-making, particularly in primary care settings where endocrine expertise may be limited. Second, the variation in performance profiles across algorithms indicates that the optimal choice depends on the specific clinical context, available computational resources, interpretability requirements, and the relative importance of sensitivity versus specificity in the target application. Third, the strong performance of relatively simple algorithms like Logistic Regression suggests that complexity should not be pursued for its own sake, and that simpler, more interpretable models may be preferable when they achieve comparable performance to black-box ensemble methods. This study has certain limitations. Model validation was performed using a single stratified train–test split, and k-fold cross-validation was not employed. Formal statistical comparisons between models, such as confidence intervals or DeLong testing for ROC-AUC, were also not conducted. Additionally, feature importance and interpretability analyses were not explored. These aspects represent important directions for future research aimed at improving robustness, clinical interpretability, and generalizability.

Conclusion

This comprehensive comparative analysis of twelve machine learning algorithms for polycystic ovary syndrome prediction has demonstrated that computational approaches, particularly gradient boosting methods, can achieve high accuracy and robust discriminative performance for this important clinical application. Gradient Boosting, XGBoost, and HistGradientBoosting emerged as the top-performing algorithms with accuracy of 92.66%, while maintaining balanced sensitivity and specificity profiles suitablefor clinical decision support. The study revealed important trade-offs between different algorithms, with Random Forest exhibiting exceptional specificity, Naive Bayes achieving the highest sensitivity, and Logistic Regression providing a compelling combination of performance and interpretability. The marked variation in performance across algorithms underscores the importance of systematic evaluation and careful consideration of specific clinical requirements when selecting approaches for healthcare applications. These findings suggest that machine learning-based predictive tools have substantial potential to enhance early PCOS diagnosis, enable risk stratification, and support clinical decision-making, ultimately contributing to improved outcomes for women affected by this prevalent endocrine disorder. Future research should focus on external validation across diverse populations, investigation of feature importance and interpretability, and prospective evaluation of these algorithms within real-world clinical workflows to fully realize their potential for improving women's health.

Competing interests

The authors declare that they have no competing interests.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics approval and consent to participate

This study did not involve any new data collection, recruitment of participants, or acquisition of biological samples. The machine learning analysis was conducted entirely using previously collected, de-identified data that is publicly available.

Availability of data and materials

This data is collected from 10 different hospitals across Kerala, India. Among the two different datasets available in kaggle the PCOS_data_without_infertility.xlsx was used to train the model.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used AI tools in order to reformulate some sentences. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT Author Contribution Statement

Usha Adiga: Conceptualization; Methodology; Supervision; Validation; Writing – Review & Editing. Vasishta Sampara: Data Curation; Software; Formal Analysis; Methodology; Visualization; Writing – Original Draft. Pedda Reddemma P.: Investigation; Resources; Data Curation; Writing – Review & Editing. Supriya P.: Data Curation; Investigation; Project Administration. Sireesha Kanchi: Software; Formal Analysis; Validation; Visualization; Writing – Review & Editing. Kasala Farzia: Writing – Original Draft; Literature Review; Editing; Project Coordination.

All authors read and approved the final manuscript. The corresponding author confirms that all contributions listed are accurate and agreed upon by all authors.

References

1. Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod. 2016;31(12):2841-2855.
2. Teede HJ, Misso ML, Costello MF, Dokras A, Laven J, Moran L, et al. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Hum Reprod. 2018;33(9):1602-1618.
3. Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nat Rev Endocrinol. 2018;14(5):270-284.
4. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertil Steril. 2004;81(1):19-25.
5. Gibson-Helm M, Teede H, Dunaif A, Dokras A. Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. J Clin Endocrinol Metab. 2017;102(2):604-612.
6. Moran LJ, Misso ML, Wild RA, Norman RJ. Impaired glucose tolerance, type 2 diabetes and metabolic syndrome in polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod Update. 2010;16(6):347-363.
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358.
8. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-1219.
9. Denny A, Raj A, Ashok A, Ram CM, George R. I-HOPE: detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE; 2019. p. 673-678.
10. Bharati S, Podder P, Mondal MRH. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: 2020 IEEE Region 10 Symposium (TENSYMP). IEEE; 2020. p. 1486-1489.
11. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
12. Norman RJ, Dewailly D, Legro RS, Hickey TE. Polycystic ovary syndrome. Lancet. 2007;370(9588):685-697.
13. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785-794.
14. Lizneva D, Suturina L, Walker W, Brakta S, Gavrilova-Jordan L, Azziz R. Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertil Steril. 2016;106(1):6-15.
15. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104-116.
16. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. New York: ACM; 2005. p. 625-632.
17. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189-1232.
18. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
19. Cutler A, Cutler DR, Stevens JR. Random forests. In: Zhang C, Ma Y, editors. Ensemble machine learning: methods and applications. New York: Springer; 2012. p. 157-175.
20. Azziz R, Carmina E, Chen Z, Dunaif A, Laven JS, Legro RS, et al. Polycystic ovary syndrome. Nat Rev Dis Primers. 2016;2:16057.
21. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken: John Wiley & Sons; 2013.
22. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.
23. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-215.
24. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.
25. Ben-Hur A, Weston J. A user's guide to support vector machines. Methods Mol Biol. 2010;609:223-239.
26. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.
27. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29.
28. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149-153.
29. Zhang H. The optimality of naive Bayes. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference. Menlo Park: AAAI Press; 2004. p. 562-567.
30. Rish I. An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. Vol 3. Seattle: IJCAI; 2001. p. 41-46.