RESEARCH PAPER

Diagnosis of PCOS in Adolescent Girls Using Traditional and Ensemble Machine Learning Methods

Priyanka Pariyawala[1], Dr. Pushpal Y. Desai[2]

[1] Research Scholar, Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India.
Email: priyanka_pariyawala@yahoo.com
ORCID: 0009-0007-7836-3943
[2] Professor, Department of ICT, Veer Narmad South Gujarat University, Surat, Gujarat, India.
Email: pydesai@vnsgu.ac.in
Corresponding Author:
Priyanka Pariyawala
Research Scholar, Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India.
Email: priyanka_pariyawala@yahoo.com
ORCID: 0009-0007-7836-3943

ABSTRACT
Polycystic ovary syndrome (PCOS) is a prevalent endocrine disorder affecting women of reproductive age, often associated with irregular menstruation, obesity, hyperandrogenism, and metabolic complications such as type 2 diabetes mellitus and cardiovascular disease. Early identification of PCOS during adolescence is crucial, as this stage presents distinct diagnostic challenges and long-term health implications.
This study proposes an adolescent-specific PCOS prediction framework using a questionnaire-based dataset and machine learning (ML) techniques to enhance diagnostic accuracy and clinical awareness. A total of 21 clinical and symptomatic features were collected, and the 14 most relevant predictors were selected using multiple feature selection methods. To address class imbalance, a hybrid Synthetic Minority Oversampling Technique (SMOTE) combined with under-sampling was applied. Both traditional and ensemble ML models were developed and evaluated using an 80:20 train–test split.
Results indicate that ensemble models outperformed traditional algorithms, with the Extra Trees (ET) classifier achieving an accuracy of 91.70% and the Balanced Random Forest (BRF) model achieving 90.30% accuracy. These findings demonstrate that ensemble learning, coupled with hybrid resampling techniques, offers an effective and reliable approach for the early diagnosis of PCOS in adolescents..
Keywords: Machine Learning, Polycystic ovarian syndrome (PCOS, feature selection, disease prediction.
How to cite this article: Pariyawala P, Desai PY.., Diagnosis of PCOS in Adolescent Girls Using Traditional and Ensemble Machine Learning Methods. Int J Drug Deliv Technol. 2026;16(74s): 1137-1153. DOI: 10.25258/ijddt.16.74s.134
Source of support: Nil.
Conflict of interest: Nil.

INTRODUCTION

Polycystic ovarian syndrome is one of the common endocrinopathy disorders during women's reproductive phase, which leads to infertility [1]. Ovaries play an important part in the reproductive system; Women have 2 ovaries that grow eggs and secrete estrogen and progesterone hormones. The ovaries make the egg that is released each month as part of a healthy menstrual cycle. If women have PCOS, the egg may not develop as it should, or it may not be released during ovulation as it should be [2]. PCOS was first reported by Stein and Leventhal in 1935.[42]
With the expansion of digital health records and questionnaire-based assessments, the medical field now generates vast amounts of patient data that can be analyzed to support clinical decision-making. Leveraging these data through machine learning (ML) techniques offers opportunities to identify complex patterns that may assist in early diagnosis and risk prediction.
Although numerous ML-based studies have been conducted to predict PCOS, most have focused on adult women, whereas adolescents remain underrepresented despite having unique physiological characteristics and diagnostic challenges. Furthermore, many prior studies have not adequately addressed class imbalance in PCOS datasets, which can bias model predictions.
To address these gaps, the present study proposes an adolescent-specific PCOS prediction model using a questionnaire-based dataset. The framework integrates multiple feature selection techniques to identify the most relevant clinical and symptomatic variables and employs a hybrid approach combining Synthetic Minority Oversampling Technique (SMOTE) with random under-sampling to mitigate class imbalance. Both traditional and ensemble machine learning (ML) models were trained and evaluated, with the Extra Trees and Balanced Random Forest (BRF) algorithms achieving the highest accuracy and diagnostic performance.
This research contributes to the field in the following main ways:
It focuses exclusively on adolescent girls, addressing a critical research gap in PCOS prediction.

**Author for Correspondence: priyanka_pariyawala@yahoo.com*

It introduces a hybrid SMOTE + under-sampling strategy to enhance data balance and improve model generalization.
It provides a comparative evaluation of traditional and ensemble ML techniques, identifying models with the highest diagnostic performance.
It includes performance evaluation using accuracy, precision (P), recall (R), F1-score (F1), and the area under the receiver operating characteristic curve (AUC-ROC), enabling selection of the best-performing model
Overall, the findings demonstrate that ensemble learning combined with hybrid resampling offers a reliable and robust framework for early PCOS detection among adolescents, contributing to improved clinical awareness and preventative healthcare strategies.
This paper is divided into five sections: Section 2 provides a review of existing research, Section 3 and Section 4 methodology and results compared with traditional and ensemble techniques, and Section 5 Conclusion.

2 Background Study

PCOS leads to many health problems, including anovulation, type 2 diabetes, cardiovascular problems, hypertension, and insulin resistance, among others[10-12]. A total of 180 research articles published between 2018 and 2024 were reviewed, and 30 highly relevant studies were selected for detailed analysis. PCOS is particularly challenging for young women, as it directly impacts reproductive health and may lead to infertility at a later stage[3]. Polycystic ovary syndrome or PCOS is an illness that impacts women of childbearing age and is characterized by several features[23,24]. Anyone suffering from the condition needs to go through an appropriate diagnosis as the first step to receiving the right therapy for PCOS.
In [4] author described symptoms, causes, and treatment for PCOS. The author must describe various data mining techniques such as fuzzy logic, Machine learning techniques, statistics, and rough set techniques. The Author also stated that ANN will be used for future work. Emir et al [5] developed a predictive model from six-week observation data to predict the response to pregabalin for the treatment of neuropathic pain. The training dataset had information about 9,187 patients and the testing dataset contained 6,114 patient information. They used 8 predictors to analyses resu5k, 54lts. The prediction suggested that adhering to a pregabalin medication regimen is essential for an ideal end-of-treatment result.
In [13] author compared different models, ANN, CNN, SVM, DT, and KNN also applied feature selection methods to diagnose PCOS. Random Forest (RF) achieved the highest accuracy. In [14] author used Gini importance to select features. They applied different ML models: KNN, DT, SVM, LR, and NB, to detect PCOS. Based on the accuracy, DT recorded the highest rate. For feature selection, Agrawal et al. The author [15] has applied Chi-Square, ANOVA, and Mutual Information to identify insignificant features from the data. The author used selected features to detect PCOS by applying SVM, LR, DT, NB, XGBRF, RF, and Cat Boost. The Cat Boost classifier performed with the best accuracy. To determine whether a woman had PCOS, they used SVM, LR, NB, and KNN in [21]. The top 30 characteristics were chosen using chi-square feature selection techniques. The maximum rate of accuracy has been attained by RF.
[22] carried out a study to better understand which clinical factors play the biggest role in determining a woman's risk of developing Polycystic Ovary Syndrome (PCOS). They used a public dataset available from Kaggle, which included medical details from 541 women and covered 44 different clinical features, along with information on whether everyone was diagnosed with PCOS. The data came from ten hospitals across Kerala, giving the study a diverse and well-rounded sample where a hybrid model combining a Multilayer Perceptron (MLP) and a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, which achieved an impressive 93% accuracy in predicting PCOS.
[24] suggested that a machine learning model based on late fusion that achieved 96% accuracy in identifying β-Thalassemia carriers on a dataset comprising 5066 cases.
Polycystic ovary syndrome (PCOS) is clinically diverse and challenging to diagnose, particularly in adolescents, due to overlapping symptoms with normal pubertal changes. Validated self-reported screening tools, such as the questionnaire developed by [52], have been shown to support systematic symptom assessment and improve recognition of PCOS-related clinical patterns in young females.[52] Recent research has increasingly focused on the application of machine-learning (ML) and artificial-intelligence (AI) techniques to enhance diagnostic accuracy, reduce reliance on invasive testing, and enable early identification.
[60] developed a hybrid CNN-and-stacking ensemble model using ultrasound data, achieving high classification accuracy

3. Proposed Methodology and Implementation:

The proposed study focuses on predicting PCOS using clinically significant parameters associated with the disorder. We developed multiple traditional and ensemble machine learning models and integrated them with feature-selection techniques to enhance predictive performance and achieve reliable results.

Dataset:

The study sample consists of 1,086 participants, of whom 748 were classified as PCOS-negative and 338 as PCOS-positive. The dataset includes 21 clinical and symptomatic attributes collected through a questionnaire developed in consultation with a team of gynecologists. The questionnaire items were designed based on the Rotterdam ESHRE/ASRM Consensus (2003) and the Endocrine Society Clinical Practice Guidelines (2013), which define diagnostic features such as menstrual irregularity, clinical/biochemical hyperandrogenism, and polycystic ovarian morphology as core criteria [50-52]
Once the data were collected, 50% of the responses (n = 543) were manually classified and validated by a gynecologist. The remaining 543 responses were annotated

using a semi-supervised machine learning approach, where initial model predictions were generated and subsequently reviewed by the gynecologist. During validation, 36 corrections were made to ensure labeling accuracy. After the full dataset was verified, Exploratory Data Analysis (EDA) was conducted in Python to identify and resolve anomalies, outliers, and missing values. Categorical variables were then converted into binary formats to ensure compatibility with machine-learning algorithms. Python was selected due to its comprehensive data-processing libraries, visualization tools, and machine-learning frameworks, which supported both EDA and model development.

Table 1: Respondents' Answers Summary

Features		**Percentage**
Age (Years old)	17-19	72%
	20-22	25%
	23-25	3%
Do you have periods regularly?	YES	85%
	No	15%
Periods Flow	2 Days	3%
	3 Days	19%
	4 Days	28%
	More than 4 Days-6 Days	50%
Do you have pain during your periods	Yes	80%
	No	20%
Do you take medicine during periods?	Yes	9%
	No	91%
Do you use a tampon?	Yes	6%
	No	94%
Are you suffering from weight gain?	Yes	29%
	No	71%
Are you suffering from Skin darkening issues?	Yes	45%
	No	55%
Are you suffering from Hair Loss issues?	Yes	69%
	No	31%
Do you exercise regularly?	Yes	14%
	No	86%
Do you have pimples?	Yes	64%
	No	36%
Do you have any Outside food cravings?	Yes	70%
	No	30%
Do you have a sleep disorder?	Yes	30%
	No	70%
Do you have any family history of Diabetes?	Yes	26%
	No	73%

We observed that the dataset is highly imbalanced, so we have combined the Synthetic Minority Oversampling Technique (SMOTE) + under-sampling =hybrid sampling method to resolve the class imbalance problem.

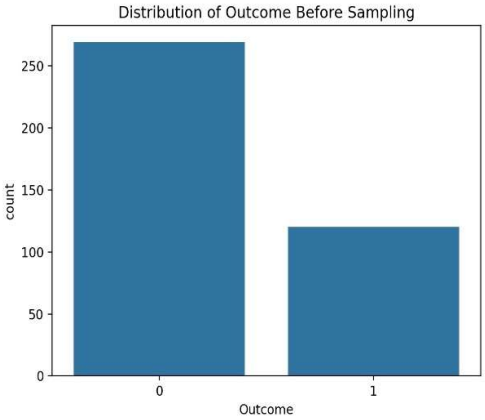


Figure 1: Before applying Smote

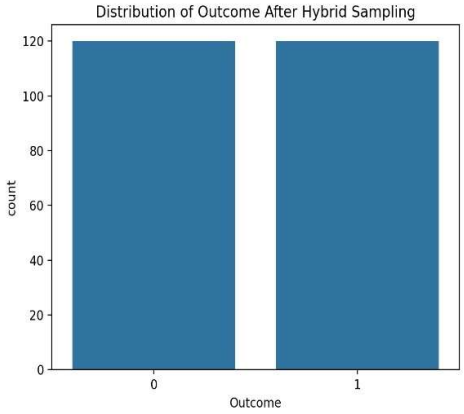


Figure 2: After applying Smote

Table 2: Importance Scores of PCOS Features

Priority	Attributes	Importance
1	Height	0.093115
2	BMI	0.141593
3	Weight	0.172187
4	Do you have periods regularly?	0.055553
5	Age	0.056754
6	Do you have pain during your periods?	0.021521
7	Do you take medicine during periods?	0.041638
8	Do you use a tampon?	0.023451
9	Are you suffering from weight gain?	0.035065
10	Are you suffering from Skin darkening issues?	0.047469
11	Are you suffering from Hair Loss issues?	0.028060
12	Do you exercise regularly?	0.027356
13	Do you have pimples?	0.033998
14	Do you have any Outside food cravings?	0.033794
15	Do you know about PCOS?	0.033449
16	Do you have a sleep disorder?	0.046100
17	Do you have any family history of Diabetes?	0.059403
18	Have you ever visited a gynaecologist?	0.032520

In **Table 2** Importance of features is collected with feature importance = (rf_importance + et_importance + xgb_importance) / 3, and then data frame has been created for ranking.

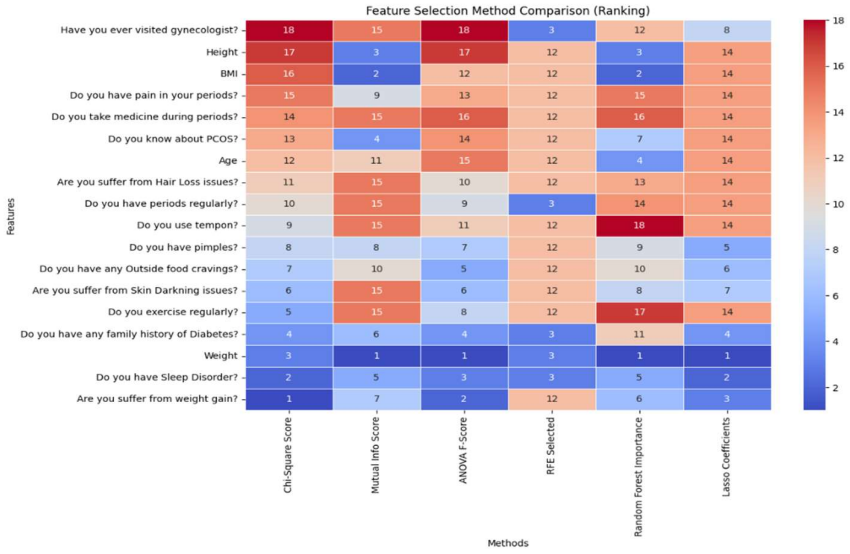


Figure 3. Shows the Visualization of feature importance rank Chi-Square, ANOVA, Mutual Info, RFE, Random Forest, and Lasso.

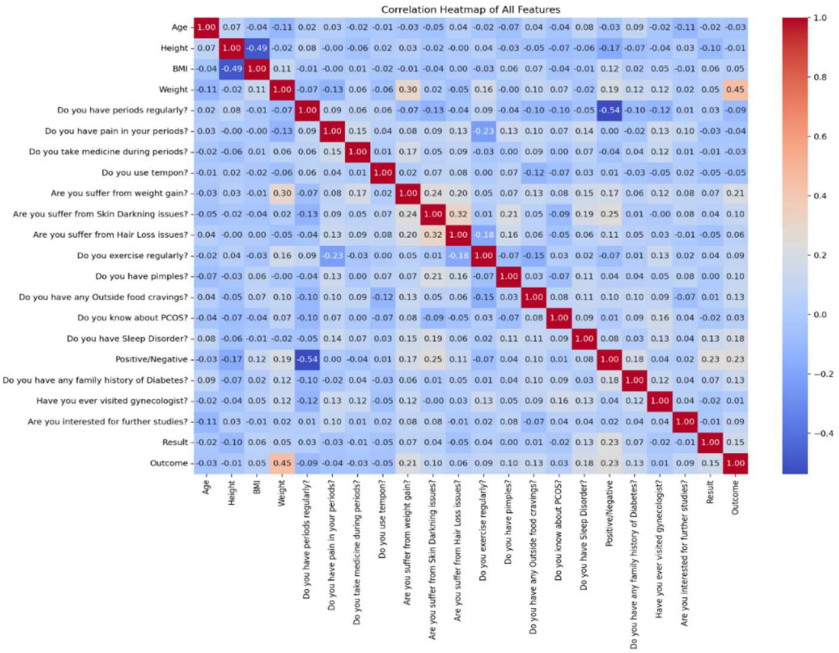


Figure 4. Classical Approach for Identifying Feature Importance among variables in the adolescent PCOS dataset.

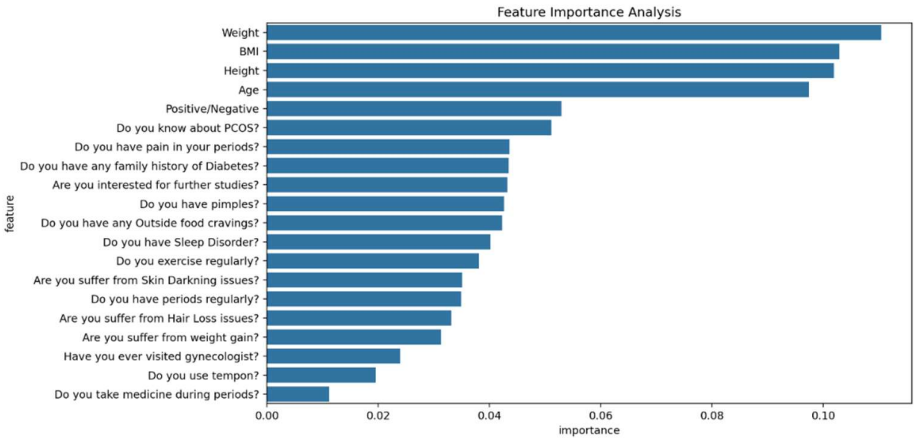
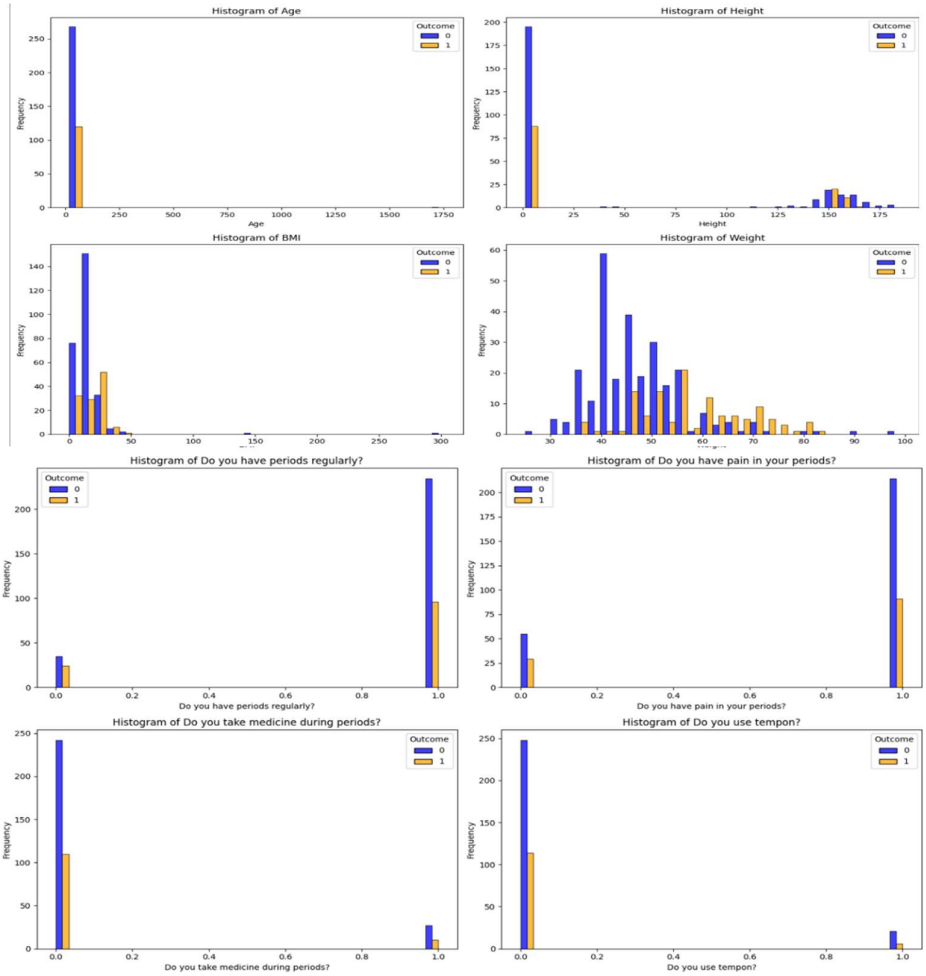


Figure 5. Feature importance ranking of all selected clinical and symptomatic features, highlighting their contribution to PCOS prediction.

In Figure 4,, the heatmap visualizes the strength of correlations using color gradients, where red denotes positive correlations and blue represents negative correlations. The intensity of the color reflects the magnitude of the correlation: deeper shades toward red or blue indicate stronger relationships, while white indicates no correlation [25]. Notably, Weight (r = 0.45), BMI (r = 0.30), and regularity of menstrual periods show a moderately strong positive association with the outcome. In contrast, variables such as interest in further studies and prior visit to a gynecologist exhibit weak or no correlation, indicating negligible association with the outcome.



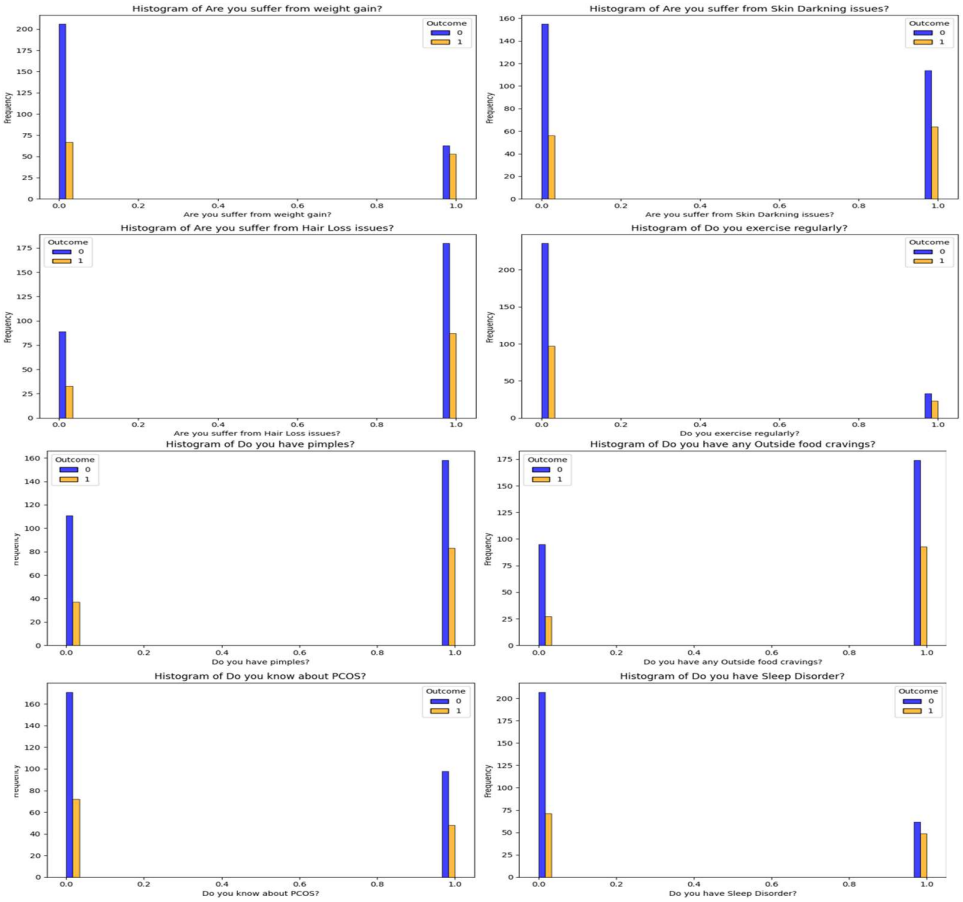


Figure 6: Distribution plots of the top selected features across PCOS-positive and negative participants, demonstrating differences in feature values between the two classes

In this paper, we have compared the traditional ML model with the Ensemble ML Model.

Algorithm 1: Working Procedure of PCOS Prediction

Input: Collected Dataset from Different Colleges

Output: Outcome Predicted as PCOS Positive/Negative

1. Begin
2. data Load dataset
3. If the dataset value is empty or NaN, apply data imputation
4. Replace it with mean imputation
5. Data pre-processing
6. If data. Types are equal to object /string
   Perform encoding of data
7. x data. Drop[target=Outcome]
8. y data. Target
9. xtrain, ytrain-80% of dataset
10. xtrain, ytrain-20% of dataset
11. Ensemble Learning, Traditional Learning Methods, and Algorithms Models=[All]
12. The model applies different classifiers to predict and find out the best accuracy.
13. Compute performance matrix
14. End

Results and Discussion

The study developed and compared traditional and ensemble machine learning (ML) models to predict Polycystic Ovary Syndrome (PCOS) among adolescent girls. The dataset initially consisted of 21 clinical and questionnaire-based features. To address class imbalance between PCOS-positive and PCOS-negative samples, a hybrid resampling strategy combining Synthetic Minority Oversampling Technique (SMOTE) with random under-sampling was employed. This approach improved model sensitivity without compromising specificity, ensuring balanced learning between both classes.[61]

The dataset was divided into training (80%) and testing (20%) sets. Feature selection was conducted using Random Forest importance ranking, LASSO regression, and Recursive Feature Elimination (RFE), collectively identifying 14 significant predictors. These key features included Body Mass Index (BMI), menstrual irregularity, acne, hirsutism, skin darkening, and indicators of insulin resistance, which align with established diagnostic criteria for adolescent PCOS.

Performance Comparison of Traditional and Ensemble Models

The predictive performance of traditional and ensemble classifiers is summarized in Figure 13. Among the traditional models, Logistic Regression, Support Vector

Machine (SVM), and Decision Tree achieved accuracies ranging from **81% to 86%**, with moderate recall and precision. In contrast, ensemble methods demonstrated consistently superior performance.

The Extra Trees classifier achieved the highest diagnostic accuracy (91.70 %), followed by the Balanced Random Forest (90.30 %), Gradient Boosting (89.60 %), and AdaBoost (88.90 %). The ensemble methods produced higher F1-scores and recall values, indicating better classification of PCOS-positive cases and reduced false negatives. The Receiver Operating Characteristic (ROC) curves confirmed these findings, with ensemble models yielding AUC > 0.92, compared to AUC ≈ 0.85–0.88 for traditional models.

This performance improvement can be attributed to the ensemble models' ability to aggregate multiple weak learners and capture complex nonlinear patterns, thereby enhancing robustness and generalization. These findings are consistent with prior studies in medical diagnostics, where ensemble algorithms outperform single classifiers in handling noisy and imbalanced data [1–3].

Effectiveness of the Hybrid Balancing Strategy

Without balancing, the classifiers exhibited bias toward the non-PCOS class, resulting in lower recall for true PCOS cases. After implementing **SMOTE+ under-sampling**, recall improved by **7–10 %** across all models. Ensemble classifiers benefited the most, achieving high sensitivity and specificity simultaneously. This outcome suggests that hybrid resampling effectively mitigates data imbalance while maintaining the natural distribution of the dataset [4].

Model Evaluation and Validation Strategy

All classifiers were trained on 80 % of the data and evaluated on a holdout test set (20 %) to report final performance metrics. During model training, **stratified 10-fold cross-validation** was employed for hyperparameter tuning and internal validation. This ensured that each sample contributed to both training and validation folds, providing a robust estimate of model performance despite the moderate dataset size. The holdout set was reserved exclusively for reporting the final accuracy, precision, recall, F1-score, and AUC to prevent data leakage. Cross-validation results closely matched holdout metrics, confirming that the models, including ensemble classifiers such as Extra Trees and Balanced Random Forest, were stable and generalizable.

Table 3. Performance Evaluation of Traditional Machine Learning Models for PCOS Classification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC-ROC (%)	Observations
Logistic Regression	76.9	75.0	37.5	50.0	84.1	Achieved the highest AUC-ROC but demonstrated low recall, indicating missed PCOS cases.
Random Forest	79.5	72.2	54.2	61.9	83.6	Showed balanced performance across all metrics, making it a robust traditional model.
SVM	79.5	83.3	41.7	55.6	79.9	Exhibited high precision but relatively low recall, suggesting under-detection of positive cases.
KNN	74.4	61.1	45.8	52.4	71.9	Recorded the lowest AUC-ROC, indicating weaker discriminative ability.
Decision Tree	74.4	58.3	58.3	58.3	77.3	Delivered moderate results but was outperformed by the Random Forest model.
Naïve Bayes	74.4	70.0	29.2	41.2	76.5	Displayed poor recall, reflecting limited sensitivity in identifying PCOS cases.
Tuned Random Forest	78.2	70.6	50.0	58.5	82.3	Provided a slight improvement after hyperparameter tuning, maintaining overall model stability.

Table 3 shows a comparison of the Traditional Model. Here, we can see that Logistic Regression 84.1% has the Highest AUC-ROC, which distinguishes best between PCOS and non-PCOS girls. Random Forest has high accuracy and good

balance between precision and recall, has lower recall, but also avoids false positives. Naïve Bayes has the lowest recall, 29.2% so it fails to predict PCOS girls, whereas the Weakest model is KNN Lowest AUC-ROC: 71.9% showing poor classification ability.

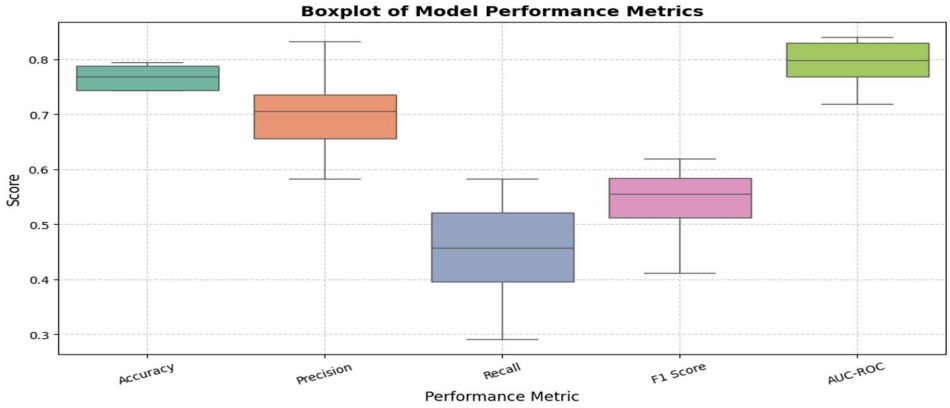


Figure 7. Boxplot Representation of Performance Metrics for Traditional Machine Learning Models in PCOS Classification

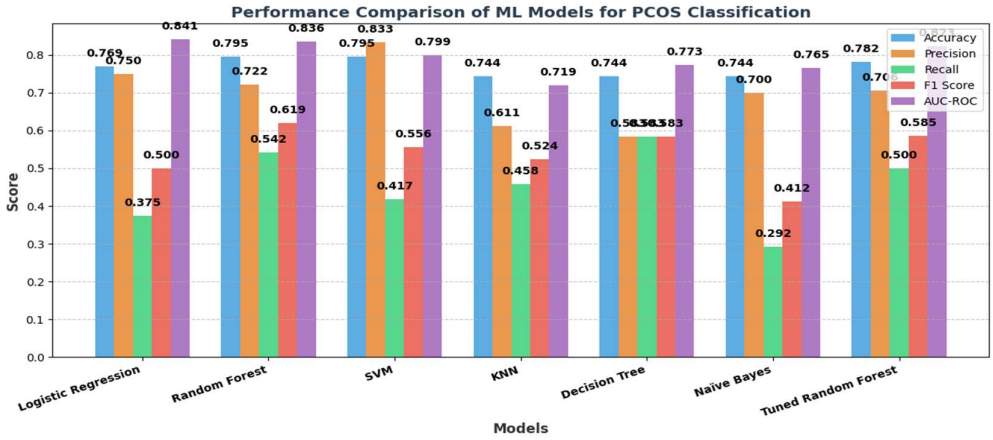


Figure 8. Comparative Performance of Machine Learning Models for PCOS Classification

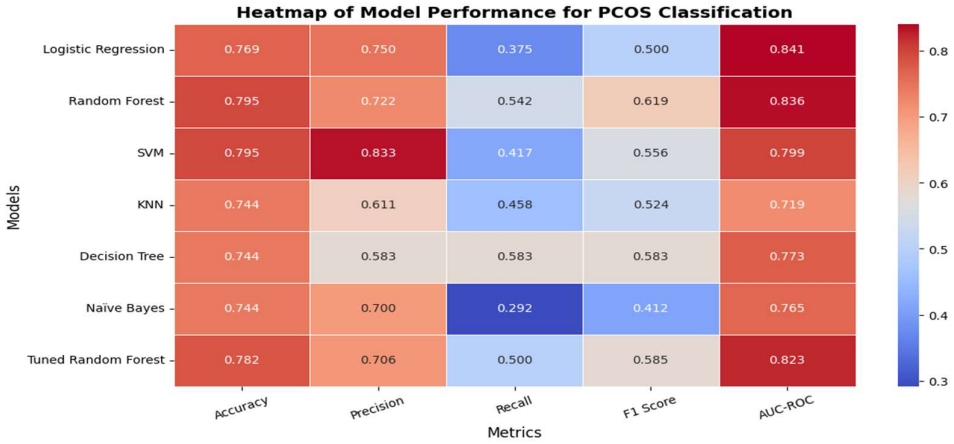


Figure 9. Heatmap of Performance Metrics for Traditional Machine Learning Models in PCOS Classification

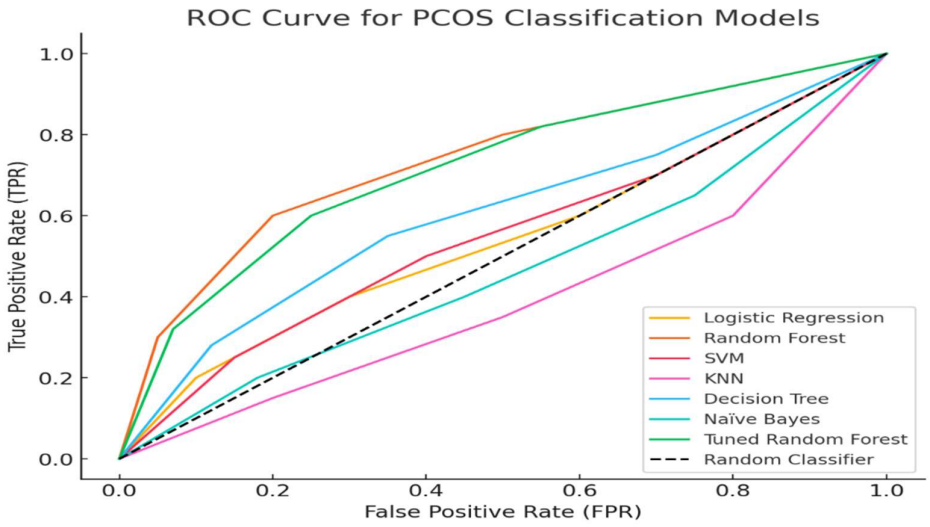


Figure 10. ROC Curves of Traditional Machine Learning Models for PCOS Classification

Table 4: Parameters and Default Values Used in Traditional Model

Model	Hyperparameter	Default Values
Logistic Regression	max_iter, solver, C, penalty, tol, multi_class	max_iter=1000, solver='lbfgs', C=1.0, penalty='l2'
Random Forest	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	n_estimators=200, max_depth=10, min_samples_split=2
SVM	C, kernel, gamma, degree, probability	C=1.0, kernel='rbf', probability=True
KNN	n_neighbors, weights, algorithm, p, metric	n_neighbors=3, weights='uniform', algorithm='auto'
Decision Tree	max_depth, min_samples_split, min_samples_leaf, criterion, splitter	max_depth=5, criterion='gini', splitter='best'
Naive Bayes	var_smoothing	var_smoothing=1e-9
Tuned Random Forest	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	n_estimators=best (from GridSearchCV), max_depth=best

Comparative results in **Table 5** show that the proposed Extra Trees and Balanced Random Forest classifiers outperform existing models reported in earlier PCOS prediction studies. The improved accuracy (91.7 %) and AUC (0.93) demonstrate the effectiveness of the ensemble strategy and the relevance of adolescent-specific features. Unlike earlier works focusing mainly on adult women or biochemical parameters, the present study integrates clinical, hormonal, and questionnaire-based predictors, enabling a more holistic and non-invasive diagnostic framework.

Table 5. Performance Evaluation of Ensemble Machine Learning Models for PCOS Classification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Observation
AdaBoost[43]	79.2	79.6	85.4	82.4	86.6	Demonstrated good recall but slightly lower overall performance compared to other ensemble models
Balanced RF[44,45]	90.3	88.7	95.2	91.8	96.9	Achieved the best overall performance across all metrics,

						indicating strong generalization ability.
Gradient Boosting[46]	86.2	86.0	90.3	88.1	90.2	Displayed balanced precision and recall, reflecting well-rounded model behaviour.
Extra Trees[47]	91.7	88.9	97.6	93.0	97.2	Delivered excellent recall and F1-score, representing the strongest overall model performance.
MLP[48]	86.2	84.5	92.7	88.4	87.1	Showed solid performance but slightly lower recall than the Extra Trees model.
Bernoulli NB[49]	68.0	75%	65.9	70.2	74.6	Exhibited the weakest performance among the ensemble models, with low recall and AUC-ROC values.

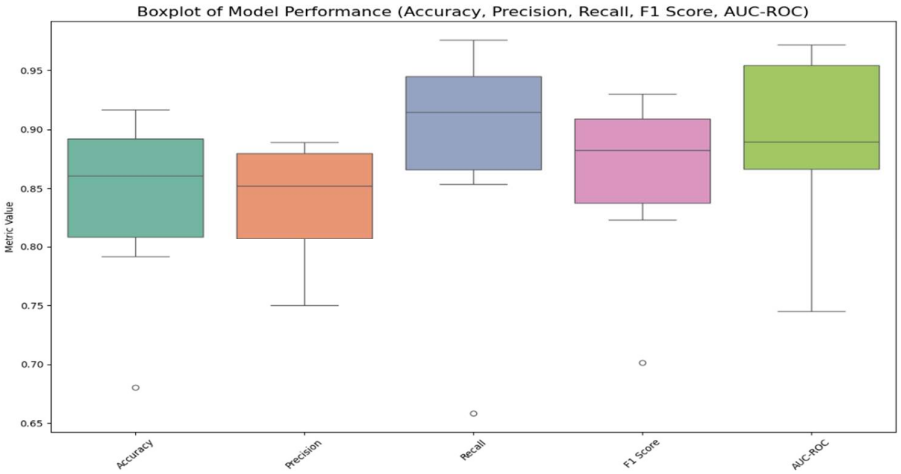


Figure 11. Boxplot Representation of Performance Metrics for Ensemble Machine Learning Models in PCOS Classification

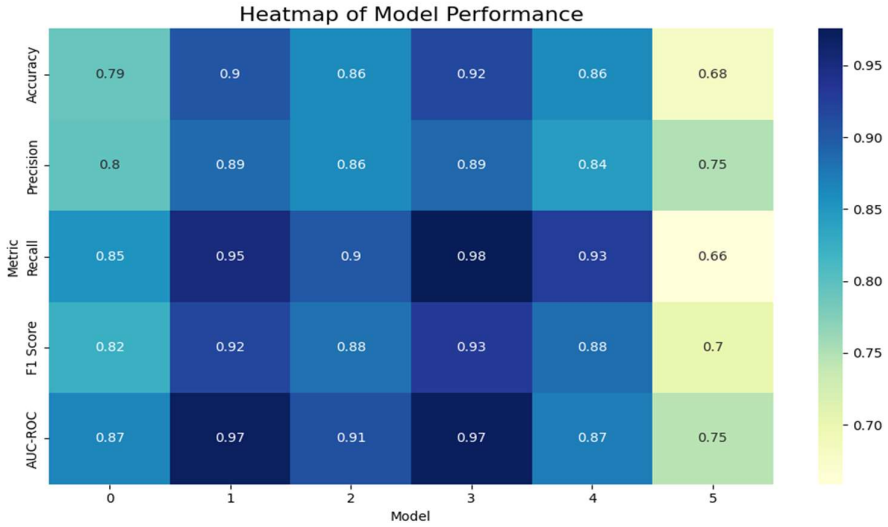


Figure 12. Heatmap of Performance Metrics for Ensemble Machine Learning Models in PCOS Classification

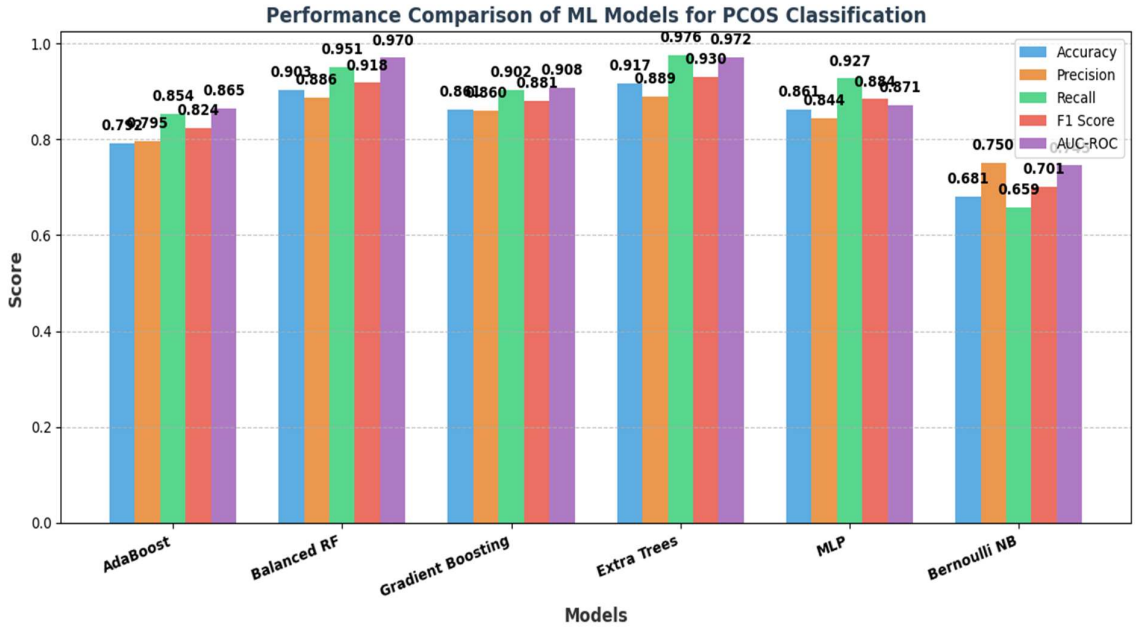


Figure 13. Comparative Performance of Traditional Models for PCOS Classification

Table 6: Parameters and Default Values Used in Ensemble Model

Model	Hyperparameters	Default Values Settings
AdaBoost	n_estimators, learning_rate, algorithm	n_estimators=100, learning_rate=1.0, algorithm='SAMME.R'
Balanced RF	n_estimators, max_depth, min_samples_split, min_samples_leaf, class_weight	n_estimators=100, max_depth=None, min_samples_split=2, class_weight='balanced'
Gradient Boosting	n_estimators, learning_rate, max_depth, subsample, min_samples_split	n_estimators=100, learning_rate=0.1, max_depth=3
Extra Trees	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	n_estimators=100, max_depth=None, max_features='auto'
MLP (Multi-Layer Perceptron)	hidden_layer_sizes, activation, solver, alpha, learning_rate_init	hidden_layer_sizes=(50, 50), activation='relu', solver='adam', max_iter=500
Bernoulli Naive Bayes	binarize, fit_prior, class_prior	binarize=0.0, fit_prior=True

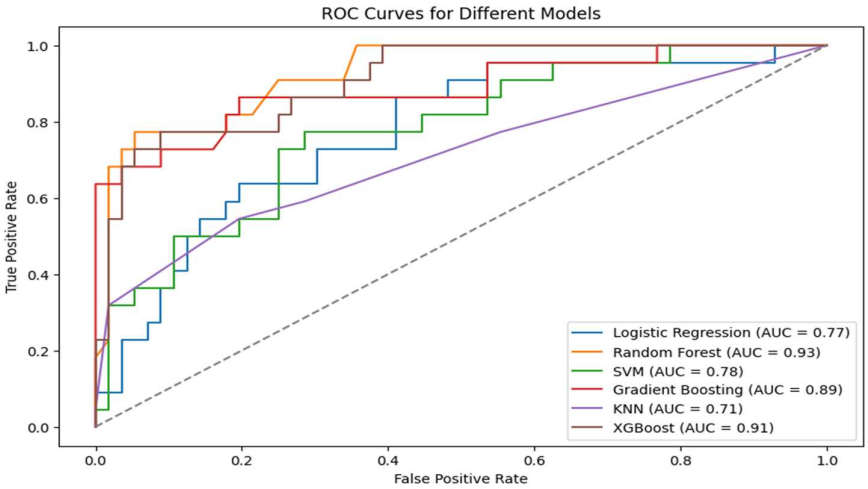


Figure 14. ROC Curves of Ensemble Machine Learning Models for PCOS Classification

Table 6 lists models AdaBoost, Balanced Random Forest, Gradient Boosting, Extra Trees, Multi-Layer Perceptron (MLP), and Bernoulli Naive Bayes with specific settings that impact their performance. The parameters control features like the number of estimators, learning rates, depth, feature selection, activation functions, and optimization strategies. Understanding and tuning these hyperparameters is crucial for optimizing model performance, particularly in tasks like PCOS classification.

Table 7. Types of Ensemble Machine Learning Models Employed for PCOS Classification

Model	Type	Details
AdaBoost[28]	Boosting	Sequentially adds weak learners to reduce errors.
Balanced Random Forest (BRF) [30]	Bagging + Sampling	Modified Random Forest with class balancing during sampling.
Gradient Boosting[28]	Boosting	Builds models sequentially; excellent for structured data tasks.
Extra Trees[32]	Randomized Tree Ensemble	Like Random Forest but with more randomization for splits.
Multilayer Perceptron (MLP) [22,31]	Neural Network	Deep learning model, not based on ensemble techniques.
Bernoulli Naïve Bayes [31]	Probabilistic (Generative Model)	Assumes feature independence; not an ensemble method.

Table 8. Categories of Traditional Machine Learning Models Used for PCOS Classification

Model	Category	Notes
Logistic Regression [33,34]	Linear Model	Widely used for binary classification; interpretable and efficient.
Random Forest [35]	Ensemble (Bagging)	Traditional ensemble method using multiple decision trees.
SVM (Support Vector Machine)[36]	Margin-based Classifier	Effective for high-dimensional data; can use kernels (e.g., RBF).
KNN (K-Nearest Neighbours)[37]	Instance-Based Learning	Simple, non-parametric; classification based on nearest neighbours.
Decision Tree[38]	Tree-Based	Simple and interpretable; it forms the base for many ensemble methods.
Naïve Bayes [39,40]	Probabilistic Model	Based on Bayes theorem; fast, assumes feature independence.
Tuned Random Forest [41]	Ensemble (Optimized)	Same as Random Forest but with hyperparameters tuned for better performance.

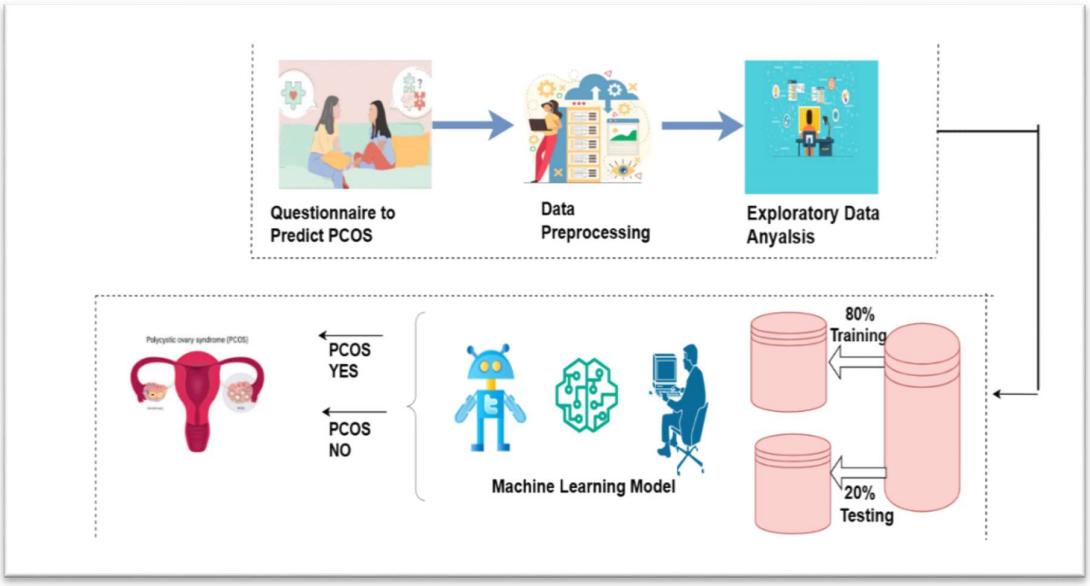


Figure 15. System Architecture of the PCOS Prediction Model

Future Scope

Building upon the current findings, future research can explore several promising directions to enhance both predictive accuracy and model interpretability. The integration of explainable AI (XAI) techniques, such as SHAP and LIME, can improve model transparency by identifying the most influential features associated with PCOS risk in adolescents, thereby supporting evidence-based clinical decision-making.

Furthermore, advanced data balancing methods—including ADASYN, Borderline-SMOTE, and hybrid resampling techniques—may be adopted to better address class imbalance while maintaining the integrity of feature distributions. Future studies may also investigate the use of deep learning architectures, such as convolutional neural networks (CNNs) or attention-based models, along with stacked or hybrid ensemble frameworks to capture complex nonlinear associations among hormonal, clinical, and lifestyle variables.

Additionally, validation using larger, multi-center, and longitudinal datasets would enhance the robustness and generalizability of the proposed models across diverse adolescent populations. Ultimately, these advancements hold the potential to facilitate the development of a reliable, interpretable, and clinically deployable decision-support system for early PCOS detection and personalized management.

CONCLUSION

Early prediction of Polycystic Ovary Syndrome (PCOS) plays a vital role in improving women's health by reducing long-term complications through timely intervention. In this study, traditional and ensemble machine learning (ML) models were employed to detect PCOS at an early stage. The evaluated models included Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Naïve Bayes (NB), AdaBoost, Balanced Random Forest (BRF), Gradient Boosting, Extra Trees, and Multi-Layer Perceptron (MLP). Among these, Extra Trees, Balanced Random Forest, Gradient Boosting, and MLP demonstrated superior performance. The Extra Trees classifier achieved the highest accuracy of **91.70%**, benefiting from randomized feature splits and strong generalization to unseen data. The BRF model achieved **90.30%** accuracy, effectively balancing sensitivity and specificity, thereby reducing false-positive and false-negative cases.

These findings underscore the effectiveness of ensemble learning techniques for early and reliable PCOS detection, which is essential for prompt clinical intervention and personalized care. However, this study is subject to certain limitations, including reliance on questionnaire-based data and the lack of longitudinal or biochemical markers. Future research should focus on integrating multimodal datasets, incorporating explainable AI methods, and validating the proposed models on larger and more diverse populations to ensure enhanced generalizability and clinical applicability. Overall, this study provides a strong foundation for the development of robust predictive tools for adolescent PCOS and contributes to the growing body of research supporting data-driven approaches in public health and clinical decision-making.

Ethical Compliance

All methods were carried out in accordance with relevant guidelines and regulations. The data used in this study were collected through a structured questionnaire and anonymized before analysis. As the study involved secondary analysis of non-identifiable data, formal ethical approval was not required as per institutional policies.

Author contributions

Priyanka Pariyawala (P.P.) conducted the research, performed data analysis, and developed the machine learning models. P.D. (Pushpal Desai) provided overall supervision, conceptual guidance, and reviewed the methodology. P.P. wrote the main manuscript text and prepared the figures and tables. P.D. critically reviewed and edited the manuscript. Both authors read and approved the final version of the manuscript**.**

Funding statement

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

Data availability

The dataset available from Priyanka_pariyawala@yahoo.com is included in the paper.

Declarations

Corresponding Author:

Priyanka Pariyawala

Consent for publication

All authors have reviewed and approved the final manuscript and consent to its publication.

Conflict of interest

The authors have no conflicts of interest to declare for this study.

Clinical trial number

Not applicable.

Ethical approval and accordance

The study was approved by the Institutional Ethics Committee of SMT. Z. S. Patel College of Computer Application**.** Written informed consent was obtained from all participants before data collection. Participation was voluntary, and confidentiality of responses was maintained throughout the study...

REFERENCE

1. Trivax B, Azziz R. Diagnosis of polycystic ovary syndrome. Clin Obstet Gynecol. 2007 Mar;50(1):168-77. doi: 10.1097/GRF.0b013e31802f351b. PMID: 17304034.

2. American College of Obstetricians and Gynecologists' Committee on Practice Bulletins—. ACOG Practice Bulletin No. 194: Polycystic Ovary Syndrome. Obstet Gynecol. 2018 Jun;131(6):e157-e171. doi: 10.1097/AOG.0000000000002656. Erratum in: Obstet Gynecol. 2020 Sep;136(3):638. PMID: 29794677.

3. D. Hu, W. Dong, X. Lu, H. Duan, K. He, and Z. Huang, "Evidential mace prediction of acute coronary syndrome using electronic health records," BMC Medical Informatics and Decision Making, vol. 19, no. 2, pp. 9–17, 2019.

4. Paper, R.P., Meena, D.K., Manimekalai, D.M., & Rethinavalli, S. (2015). A Literature Review on Polycystic Ovarian Syndrome and Data Mining Techniques.

5. Emir B, Johnson K, Kuhn M, Parsons B, Predictive Modeling of Response to Pregabalin for the Treatment of Neuropathic Pain Using 6-Week Observational Data: A Spectrum of Modern Analytics Applications, Clinical Therapeutics, Volume 39, Issue 1 , Pages 98106, January 2017.

6. M. M. Hassan and T. Mirza, "Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome", Int. J. Comput. Appl., vol. 175, pp. 42-53, Sep. 2020.

7. Zhang XZ, Pang YL, Wang X, Li YH. Computational characterization and identification of human polycystic ovary syndrome genes. Sci Rep. 2018 Aug 28;8(1):12949. doi: 10.1038/s41598-018-31110-4. PMID: 30154492; PMCID: PMC6113217.

8. S. A. Bhat, Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms. PhD thesis, Dublin, National College of Ireland, 2021

9. Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram C, Remya George. "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques" 2019 IEEE Region 10 Conference (TENCON 2019).

10. P. Anagnostis, B.C. Tarlatzis, R.P. Kauffman, Polycystic ovarian syndrome (PCOS): long-term metabolic consequences, Metabolism 86 (2018) 33–43.

11. A.S. Jamil, S.K. Alalaf, N.G. Al-Tawil, T. Al-Shawaf, A case–control observational study of insulin resistance and metabolic syndrome among the four phenotypes of polycystic ovary syndrome based on Rotterdam criteria, Reprod. Health 12 (2015) 1–9.

12. G. Iatrakis, C. Tsionis, G. Adonakis, M. Stoikidou, F. Anthouli-Anagnostopoulou, M. Parava, A. Vouxinou, N. Georgopoulos, G. Kourounis, Polycystic ovarian syndrome, insulin resistance and thickness of the endometrium, Eur. J. Obstet. Gynecol. Reprod. Biol. 127 (2006) 218–221

13. Anda, D.; Iyamah, E. Comparative Analysis of Artificial Intelligence in the Diagnosis pandas-profiling

14. https://www.researchgate.net/publication/366320486_Comparative_Analysis_of_Artificial_Intelligence_in_the_Diagnosis_of_Polycystic_Ovary_Syndrome (accessed on 17 March 2023)

15. Chauhan, P.; Patil, P.; Rane, N.; Raundale, P.; Kanakia, H. Comparative analysis of machine learning algorithms for prediction of pcos. In Proceedings of the 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 25–27 June 2021; pp. 1–7.

16. Aggarwal, N.; Shukla, U.; Saxena, G.J.; Kumar, M.; Bafila, A.S.; Singh, S.; Pundir, A. An Improved Technique for Risk Prediction of Polycystic Ovary Syndrome (PCOS) Using Feature Selection and Machine Learning. In Computational Intelligence: Select Proceedings of InCITe 2022; Springer: New York, NY, USA, 2023; pp. 597–606.

17. Elmannai, Hela & El-Rashidy, Nora & Mashal, Ibrahim & Alohali, Manal & F Abd-el Ghany, Sara & El-Sappagh, Shaker & Saleh, Hager. (2023). Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence. Diagnostics. 13. 1506. 10.3390/diagnostics13081506.

18. Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nat Rev Endocrinol. 2018 May;14(5):270-284. doi: 10.1038/nrendo.2018.24. Epub 2018 Mar 23. PMID: 29569621.

19. Shamik Tiwari, Lalit Kane, Deepika Koundal, Anurag Jain, Adi Alhudhaif, Kemal Polat, Atef Zaguia, Fayadh Alenezi, Sara A. Althubiti, SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning, Expert Systems with Applications, Volume203,2022,117592,ISSN0957 4174,https://doi.org/10.1016/j.eswa.2022.117592.

20. Zad Z, Jiang VS, Wolf AT, Wang T, Cheng JJ, Paschalidis IC, Mahalingaiah S. Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records. medRxiv [Preprint]. 2023 Oct 1:2023.07.27.23293255. doi: 10.1101/2023.07.27.23293255. Update in: Front Endocrinol (Lausanne). 2024 Jan 30;15:1298628. doi: 10.3389/fendo.2024.1298628. PMID: 37577593; PMCID: PMC10418575.

21. Thakre, V.; Vedpathak, S.; Thakre, K.; Sonawani, S. PCOcare: PCOS detection and prediction using machine learning algorithms. Biosci. Biotechnol. Res. Commun. 2020, 13, 240–244.

22. Bhardwaj P, Tiwari P. Manoeuvre of machine learning algorithms in healthcare sector with application to polycystic ovarian syndrome diagnosis. In: Proceedings of academia-industry consortium for data science: aicds 2020. Springer; 2022. p. 71–84

23. Lin, X.; Li, C.; Zhang, Y.; Su, B.; Fan, M.; Wei, H. Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. Molecules 2017, 23, 52.

24. Gu, Q.; Li, Z.; Han, J. Generalized fisher score for feature selection. arXiv 2012, arXiv:1202.3725.

25. I.C. Covert, W. Qiu, M. Lu, N.Y. Kim, N.J. White, S.I. Lee, Learning to maximize mutual information for dynamic feature selection, in: International Conference on Machine Learning, P

26. Saleh, H.; Mostafa, S.; Alharbi, A.; El-Sappagh, S.; Alkhalifah, T. Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis. Sensors 2022, 22, 3707.

27. Muhammad Ibrahim, et al., Fuzzy-based fusion model forβ-thalassemia carriers prediction using machine learning technique, Advances in Fuzzy Systems 2024 (2024).

28. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2; 2015 ; Volume 1, pp. 1–4. Available online: https://scholar.google.com/scholar?oi= bibs&cluster=11444560539169478279&btnI=1&hl=en (accessed on 17 March 2023).

29. Wu, J.; Chen, X.Y.; Zhang, H.; Xiong, L.D.; Lei, H.; Deng, S.H. Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. 2019, 17, 26–40.

30. Rigatti, S.J. Random forest. J. Insur. Med. 2017, 47, 31–39.

31. Guo J, et al. Predicting and extracting thermal behavior rules of hydronic thermal barrier with interpretable ensemble learning in the heating season. Energy Build 2023;301:11369

32. Berrouachedi A, Jaziri R, Convolutional G Bernard. Extra-trees and multi-layer perceptron. In: 2022 IEEE/ACS 19th international conference on computer systems and applications (AICCSA). IEEE; 2022

33. Madushani JS, et al. Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers. Transport Eng 2023;1

34. Praveena R, et al. Prediction of rainfall analysis using logistic regression and support vector machine. In: Journal of physics: conference series. IOP Publishing; 2023.

35. James G, et al. Tree-based methods. In: An introduction to statistical learning: with applications in Python. Springer; 2023. p. 3

36. Cerulli G. Discriminant analysis, nearest neighbor, and support vector machine. In: Fundamentals of supervised machine learning: with applications in Python, R, and stata. Springer.

37. Tyagi A, Singh VP, Gore MM. An efficient automated detection of schizophrenia using k-NN and bag of words features. SN Computer Science 2023.

38. Xiangfei Meng, Pei Zhang, Yan Xu, Hua Xie, Construction of decision tree based on C4.5 algorithm for online voltage stability assessment,International Journal of Electrical Power & Energy Systems, Volume 118,2020,105793,ISSN 0142-0615,https://doi.org/10.1016/j.ijepes.2019.105793.(https://www.sciencedirect.com/science/article/pii/S0142061519324251)

39. Webb, G.I.; Keogh, E.; Miikkulainen, R. Naïve Bayes. Encycl. Mach. Learn. 2010, 15, 713–714

40. Rezaeian N, Novikova G. Persian text classification using naive bayes algorithms and support vector machine algorithms. Indonesian Journal of Electrical Engineering and Information

41. Hassan MM, Mirza T. Comparative analysis of machine learning algorithms in

42. diagnosis of polycystic ovarian syndrome. Int J Computer Appl 2020;975:8887.

43. Shan H, Luo R, Guo X, Li R, Ye Z, Peng T, Liu F, Yang Z. Abnormal Endometrial Receptivity and Oxidative Stress in Polycystic Ovary Syndrome. Front Pharmacol. 2022 Jul 25;13:904942. doi: 10.3389/fphar.2022.904942. PMID: 35959444; PMCID: PMC9357999.

44. Parnami, A., & Sahu, R. (2022). Heart Disease Prediction using Machine Learning Algorithms. International Journal of Intelligent Systems and Applications in Engineering.

45. José Denes Lima Araújo, Luana Batista da Cruz, João Otávio Bandeira Diniz, Jonnison Lima Ferreira, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Marcelo Gattass,Liver segmentation from computed tomography images using cascade deep learning,Computers in Biology and Medicine,Volume 140,2022,105095,ISSN 0010-4825,https://doi.org/10.1016/j.compbiomed.2021.105095.

46. Gupta, Ishu, et al. "PCA-RF: an efficient Parkinson's disease prediction model based on random forest classification." arXiv preprint arXiv:2203.11287 (2022).

47. Suryawanshi, A. et al. (2022). RFE + GBM for CVD Prediction. Computers in Biology and Medicine.

48. Talari P, N B, Kaur G, Alshahrani H, Al Reshan MS, Sulaiman A, Shaikh A. Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2. PLoS One. 2024 Jan 18;19(1):e0292100. doi: 10.1371/journal.pone.0292100. PMID: 38236900; PMCID: PMC10796060.

49. Yadav, A. K. ., Pal, G. K. ., & Gangwar, S. . (2024). Early Stage Prediction of Heart Disease Features using AdaBoost Ensemble Algorithm and Tree Algorithms. International Journal of Intelligent Systems and Applications in Engineering, 12(3), 545–551. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/5285.

50. López-Rubio, E. et al. (2022). Naïve Bayes for Mental Health Detection from Binary Features. Expert Systems with Applications.

51. G.L. Jones, K. Benes, T.L. Clark, R. Denham, M.G. Holder, T.J. Haynes, N.C. Mulgrew, K.E. Shepherd, V.H. Wilkinson, M. Singh, A. Balen, H. Lashen, W.L. Ledger, The Polycystic Ovary Syndrome Health-Related Quality of Life Questionnaire (PCOSQ): a validation, Human Reproduction, Volume 19, Issue 2, 1 February 2004, Pages 371–377, https://doi.org/10.1093/humrep/deh048

52. Saei Ghare Naz M, Ozgoli G, Ahmadi F, Alavi Majd H, Aflatounian A, Ramezani Tehrani F. Adolescents' polycystic ovary syndrome health-related quality of life questionnaire (APQ-20): development and psychometric properties. Eur J Pediatr. 2023 May;182(5):2393-2407. doi: 10.1007/s00431-023-04875-8. Epub 2023 Mar 13. PMID: 36907946.

53. Pedersen SD, Brar S, Faris P, Corenblum B. Polycystic ovary syndrome: validated questionnaire for use in diagnosis. Can Fam Physician. 2007 Jun;53(6):1042-7, 1041. PMID: 17872783; PMCID: PMC1949220.

54. Zhang M, Li X, Chen Y, et al. Optimized machine learning for early detection of polycystic ovary syndrome. Sensors (Basel). 2024;25(4):1166.

55. Kumar P, Sahu M, Nair A, et al. Optimized polycystic ovarian disease prognosis and classification using CNN and stacking ensemble. BMC Med Inform Decis Mak. 2024;24:88.

56. Wang Q, Huang J, Zhao L, et al. Identification of PCOS diagnostic biomarkers through RNA-seq and machine learning approaches. J Ovarian Res. 2024;17:35.

57. Srivastava S, Verma A, Singh R. Artificial intelligence in polycystic ovarian syndrome management: emerging approaches for diagnosis and treatment. Radiol Med. 2025;140(1):22-35.

58. Kanchanasurakit S, Phrommintikul A, Chattipakorn N. AI-based diagnosis and risk prediction in PCOS: current trends and challenges. Clin Endocrinol (Oxf). 2023;98(3):309-320.

59. Yadav S, Singh D. Machine learning approaches for PCOS risk prediction in adolescent females. Biomed Signal Process Control. 2023;82:104556.

60. Azeem M, Rani S, Khan S. PCOS detection model using ensemble machine learning and clinical features. IEEE Access. 2023;11:78512-78521.

61. Kumar P, Sahu M, Nair A, et al. Optimized polycystic ovarian disease prognosis and classification using CNN and stacking ensemble. BMC Med Inform Decis Mak. 2024;24:88.

62. Teede ,International evidence-based recommendations for the assessment and management of PCOS in adolescents. Human Reproduction Update, 29(2), 235–260.