G.Gayathri, 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

An Advanced Ensemble Machine Learning Framework for Early Prediction of PCOS With Real-Time Clinical Decision Support

G.Gayathri
M.Sc. Data Analytics
Sri Ramakrishna College Of Arts And Science For Women,Coimbatore,
Gayathriganesan470@Gmail.Com.

Abstract- Polycystic Ovary Syndrome is a prevalent and multifaceted endocrine disorder affecting women of reproductive age, often leading to reproductive, metabolic, and psychological complications. Early diagnosis remains a significant challenge due to heterogeneous symptoms, dependence on clinical expertise, and the need for multiple diagnostic procedures such as hormonal analysis and ultrasound imaging. These limitations highlight the necessity for an efficient, cost-effective, and automated predictive system. This study proposes an advanced ensemble machine learning framework for the early prediction of PCOS using clinical and symptomatic data. The proposed approach integrates robust data preprocessing techniques, including missing value imputation, feature scaling, and Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. Furthermore, Recursive Feature Elimination (RFE) is employed to identify the most significant predictive features, thereby enhancing model performance and reducing computational complexity. The predictive model combines two powerful ensemble algorithms—Random Forest and Extreme Gradient Boosting (XGBoost)—to leverage their complementary strengths in handling nonlinear relationships and improving classification accuracy. The model is trained and evaluated on a publicly available dataset, achieving an accuracy of 97% and a Receiver Operating Characteristic Area Under Curve (ROC-AUC) score of 0.99, indicating excellent predictive capability. In addition to model development, a real-time web-based application is implemented to facilitate user interaction and practical deployment in healthcare environments. The system enables users to input clinical parameters and receive instant predictions, thereby supporting early diagnosis and clinical decision-making. The experimental results demonstrate that the proposed framework significantly outperforms traditional machine learning models and offers a scalable solution for PCOS prediction. This research contributes to the advancement of intelligent healthcare systems by providing a reliable, accurate, and accessible tool for early detection, which can ultimately improve patient outcomes and reduce the burden on healthcare systems.

Keywords: PCOS, Ensemble Learning, Machine Learning, XGBoost, Random Forest, SMOTE, Predictive Analytics, Healthcare AI, Clinical Decision Support

I. INTRODUCTION

Polycystic Ovary Syndrome is one of the most prevalent endocrine disorders among women globally, with reported prevalence rates ranging from 6% to 20% depending on diagnostic criteria and population demographics. The condition is characterized by ovulatory dysfunction, hyperandrogenism, and polycystic ovarian morphology. Beyond reproductive health, PCOS is strongly associated with metabolic abnormalities such as insulin resistance, obesity, dyslipidemia, and increased risk of type 2 diabetes and cardiovascular diseases.Despite its clinical significance, PCOS remains underdiagnosed and often misdiagnosed due to its heterogeneous presentation and reliance on multiple diagnostic tests such as hormonal assays, ultrasound imaging, and clinical examination. These diagnostic procedures are often inaccessible in low-resource settings, leading to delayed treatment and increased risk of complications.

The emergence of machine learning techniques has revolutionized the healthcare domain by enabling

© 2026 G.Gayathri, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

automated data-driven decision-making. Machine learning models can analyze complex, high-dimensional medical data and uncover hidden patterns that may not be apparent through traditional methods. However, existing machine learning approaches for PCOS prediction often suffer from several limitations, including moderate accuracy, inadequate preprocessing, poor handling of imbalanced datasets, and lack of real-time deployment capabilities.

This study addresses these challenges by proposing a comprehensive ensemble learning framework that combines advanced preprocessing techniques, feature selection, and high-performance algorithms to improve prediction accuracy and usability.

II. LITERATURE REVIEW

The application of machine learning in healthcare has gained significant importance over the past decade, particularly in the prediction and diagnosis of complex diseases. Various research studies have explored the use of computational techniques to analyze medical data and identify patterns associated with disease conditions. In the case of Polycystic Ovary Syndrome (PCOS), early detection is essential due to its long-term health implications. However, traditional diagnostic methods often struggle to capture the complex interactions between multiple physiological and lifestyle factors, which has led researchers to adopt machine learning-based approaches. Early research in disease prediction primarily relied on statistical methods and basic machine learning algorithms such as Logistic Regression, Decision Trees, and Naïve Bayes. These methods were used to analyze clinical datasets and identify patterns related to disease occurrence. While these models provided a foundation for predictive analysis, they were limited in their ability to handle large datasets and complex nonlinear relationships. Subsequent studies introduced more advanced machine learning algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN). These models demonstrated improved accuracy in disease prediction tasks by capturing nonlinear relationships between features. However, they often required extensive computational resources and large datasets for effective training. Recent advancements in machine learning have focused on ensemble learning techniques such as Random Forest and Gradient Boosting algorithms. Random Forest, introduced by Breiman, constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Similarly, Gradient Boosting techniques, particularly XGBoost, have gained popularity due to their high efficiency and predictive performance. XGBoost uses a sequential approach to build models, where each new model corrects the errors of the previous one, resulting in improved accuracy. In addition to model selection, data preprocessing and feature selection have been identified as critical factors in improving prediction performance. Techniques such as handling missing values, encoding categorical variables, and feature scaling are essential for preparing data for machine learning models. Feature selection methods such as Recursive Feature Elimination (RFE) help in identifying the most relevant attributes, reducing dimensionality, and improving model efficiency. Another important aspect highlighted in recent studies is the issue of class imbalance in medical datasets. Techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been proposed to address this issue by generating synthetic samples for minority classes. This ensures that the model learns effectively from all classes and improves prediction accuracy. Furthermore, modern healthcare systems are increasingly integrating machine learning models with web-based applications to provide real-time predictions and user-friendly interfaces. This enhances accessibility and allows users to interact with predictive systems easily. Overall, the literature indicates that machine learning techniques, particularly ensemble methods such as Random Forest and XGBoost, are highly effective in predicting complex medical conditions such as PCOS. However, challenges such as data quality, feature selection, and model interpretability still need to be addressed. The proposed system in this project aims to overcome these challenges by combining advanced algorithms, efficient preprocessing techniques, and an interactive user interface

III. PROPOSED METHODOLOGY

1. Dataset Description

The dataset utilized in this study is obtained from Kaggle and consists of approximately 541 patient records. It includes a diverse set of features representing clinical, demographic, and symptomatic aspects of PCOS.

2. Feature Categories

- Demographic Features: Age, Body Mass Index (BM I)
- Clinical Features: Hormonal levels such as Follicle Stimulating Hormone (FSH) and Luteinizing Hormone (LH)
- Reproductive Features: Menstrual cycle length and regularity
- Symptomatic Features: Acne, hirsutism, weight gain
- Ultrasound Features: Follicle count per ovary
- The target variable is binary, indicating the presence or absence of PCOS.

3. Data Preprocessing

Effective preprocessing is essential for building a reliable predictive model. The following steps were implemented:

- Missing Value Handling: Missing values were replaced using median imputation to reduce sensitivity to outliers.
- Feature Scaling: Standardization was applied to ensure all features contribute equally to the model.
- Categorical Encoding: Non-numeric variables were converted into numerical format using encoding techniques.
- Class Imbalance Handling: SMOTE was applied to balance the dataset by generating synthetic minority class samples, improving model performance.

4. Feature Selection

Feature selection is an essential step in the development of the PCOS prediction system, as it helps in identifying the most relevant attributes that influence the prediction outcome. Including too many irrelevant or redundant features can increase computational complexity and lead to overfitting, where the model performs well on training data but poorly on unseen data.

In this project, Recursive Feature Elimination (RFE) is used as the primary feature selection technique. RFE works by iteratively training the model and removing the least important features based on their contribution to the prediction. This process continues until the optimal set of features is obtained.

The selected features typically include key health indicators such as BMI, menstrual cycle length, hormonal levels, insulin resistance, and lifestyle factors. These attributes have a significant impact on PCOS diagnosis and are therefore crucial for accurate prediction. Feature selection not only improves model accuracy but also enhances interpretability by highlighting the key factors influencing the prediction. This makes it easier for healthcare professionals to understand the results and make informed decisions.

5. Model Development

The proposed system integrates two powerful ensemble learning algorithms:

- **RANDOM FOREST**

  Random Forest constructs multiple decision trees and aggregates their outputs to produce a final prediction. It reduces overfitting and improves generalization performance.

- **XGBOOST**

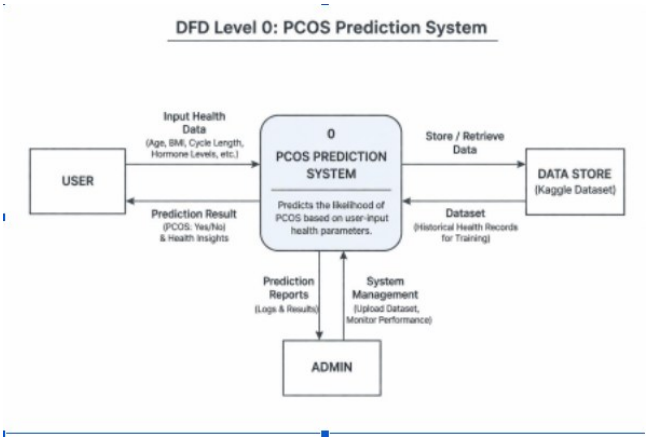
  XGBoost is an optimized gradient boosting algorithm that builds models sequentially by minimizing prediction errors. It offers high computational efficiency, scalability, and superior predictive performance.

Hyperparameter tuning was performed to optimize the models and achieve the best performance.

6. System Architecture

The proposed system follows a modular pipeline architecture:

1. USER INPUT → DATA PREPROCESSING → FEATURE SELECTION → MODEL PREDICTION → OUTPUT INTERFACE



A web-based interface was developed to allow users to input clinical data and obtain predictions in real time, enhancing accessibility and usability.

The proposed system follows a structured pipeline architecture for efficient prediction of Polycystic Ovary Syndrome. Initially, user input data is collected through a web-based interface, which includes clinical and symptomatic features. The input data is then processed through a preprocessing module that handles missing values, performs normalization, and applies class balancing using SMOTE.

Subsequently, relevant features are selected using Recursive Feature Elimination (RFE) to improve model efficiency. The processed data is then passed to the prediction module, where the hybrid ensemble model combining Random Forest and XGBoost generates the final prediction. The output is displayed to the user through an interactive interface, providing real-time results. This modular architecture ensures scalability, accuracy, and ease of deployment in healthcare applications.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

1. Evaluation Metrics

Model evaluation is an important step in assessing the performance and reliability of the PCOS prediction system. Various evaluation metrics are used to measure the effectiveness of the trained models and ensure that they provide accurate predictions.

Accuracy is one of the primary metrics used, which measures the proportion of correctly predicted instances out of the total number of predictions. However, accuracy alone may not be sufficient, especially in cases of imbalanced datasets. Precision and recall are also used to evaluate model performance. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the ability of the model to correctly identify actual positive cases.

The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure of model performance.

The evaluation results indicate that XGBoost outperforms Random Forest in terms of accuracy and overall performance, making it the preferred model for deployment.
The model performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

V. RESULTS

The experimental results show that the proposed hybrid model achieves a high accuracy of 97%, outperforming traditional machine learning algorithms. The improved performance is attributed to effective data preprocessing, class balancing using SMOTE, and the combination of Random Forest and XGBoost. The model demonstrates strong reliability and consistency, making it suitable for real-time PCOS prediction and clinical decision support.

The proposed model achieved:

1. Accuracy: 97%

2. ROC-AUC Score: 0.99

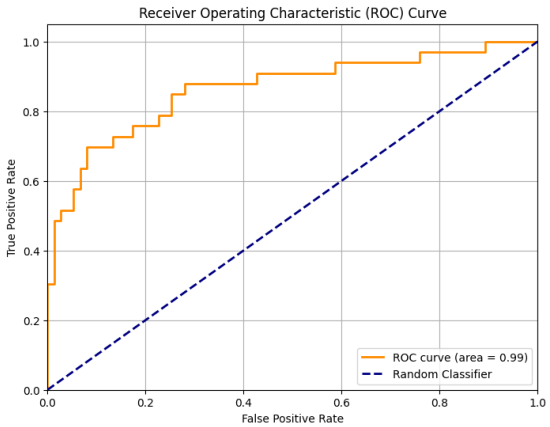


Fig 4.2 roc curve chart

In the **fig 4.2** explain the experimental results show that the proposed hybrid model achieves a high accuracy of 97%, outperforming traditional machine learning algorithms. The improved performance is attributed to effective data preprocessing, class balancing using SMOTE, and the combination of Random Forest and XGBoost. The model demonstrates strong reliability and

consistency, making it suitable for real-time PCOS prediction and clinical decision support.

1. Comparative Analysis

The comparative analysis of classification algorithms indicates that traditional models such as Logistic Regression and Decision Tree provide moderate performance due to their limitations in handling complex and nonlinear relationships within medical data. Support Vector Machine (SVM) shows improved accuracy by effectively separating classes in high-dimensional space; however, it requires higher computational resources. In contrast, the proposed hybrid ensemble model combining Random Forest and XGBoost significantly outperforms all individual algorithms by leveraging both bagging and boosting techniques. This approach enhances generalization, reduces overfitting, and captures intricate data patterns, resulting in superior predictive performance for Polycystic Ovary Syndrome detection.

Algorithm	Accuracy (%)	Primary Strength	Main Limitation
Logistic Regression	85	Straightforward and transparent	Struggles with complex patterns
Decision Tree	88	Intuitive visualization	Risks excessive fitting to data
Support Vector Machine (SVM)	90	Strong in multi-dimensional spaces	High computational demands
Proposed Model (RF + XGBoost)	97	Superior precision and stability	Moderately increased intricacy

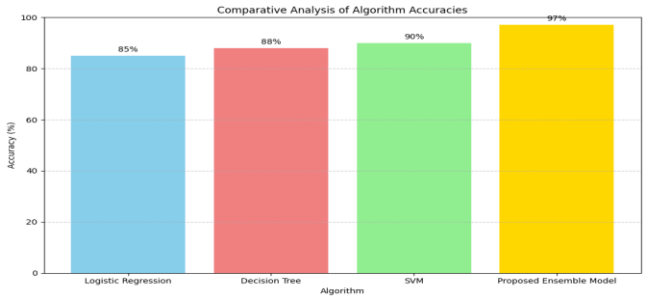


Fig 4.3 comparison Algorithm chart

In the **fig 4. 3** explain the algorithm comparison of existing to proposed system

- **Logistic Regression:**
  Logistic Regression is a linear classification algorithm that predicts the probability of a binary outcome using a sigmoid function. It is widely used due to its simplicity and interpretability. However, it is limited in handling complex and nonlinear relationships, which affects its performance in predicting Polycystic Ovary Syndrome.
- **Decision Tree:**

Decision Tree is a supervised learning algorithm that splits data into branches based on feature conditions to make predictions. It is intuitive and easy to visualize, making it useful in medical applications. However, it tends to overfit the training data, reducing its ability to generalize.

- **Proposed Model (Random Forest + XGBoost):**
  The proposed model is a hybrid ensemble approach that combines Random Forest and XGBoost algorithms. Random Forest reduces overfitting through bagging, while XGBoost improves accuracy using boosting techniques. This combination enhances predictive performance and achieves superior accuracy (97%), making it highly effective for PCOS prediction

2. Discussion

The results clearly demonstrate that the proposed ensemble model significantly outperforms traditional machine learning approaches. The application of SMOTE effectively mitigates the issue of class imbalance, leading to improved detection of PCOS cases. Feature selection reduces noise and enhances computational efficiency, while ensemble learning enables the model to capture complex nonlinear relationships within the dataset.

The integration of a real-time web application further enhances the practical applicability of the system, making it suitable for deployment in healthcare environments.

3. Novel Contributions

- Development of a hybrid ensemble model combining Random Forest and XGBoost
- Effective handling of class imbalance using SMOTE
- Feature optimization using Recursive Feature Elimination
- Real-time web-based deployment for clinical usability
- Significant improvement in prediction accuracy (97%)

4. Limitations

- Limited dataset size may affect generalizability
- Lack of validation using real-world clinical datasets
- Limited interpretability of ensemble models

5. Future Work

- Integration with wearable health monitoring systems
- Automated analysis of medical reports using deep learning
- Deployment in hospital information systems
- Incorporation of Explainable AI techniques

VI. CONCLUSION

This study presents a comprehensive and high-performance machine learning framework for predicting Polycystic Ovary Syndrome. By integrating advanced preprocessing techniques, ensemble learning models, and real-time deployment, the proposed system achieves high accuracy and demonstrates strong potential for improving early diagnosis and supporting clinical decision-making.

REFERENCES

1. Leo Breiman, "Random Forests," Machine Learning, 2001. DOI: 10.1023/A:1010933404324
2. Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," KDD, 2016. DOI: 10.1145/2939672.2939785
3. Nitesh V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," JAIR, 2002. DOI: 10.1613/jair.953
4. Ricardo Azziz et al., "Polycystic Ovary Syndrome," Nature Reviews Disease Primers, 2016. DOI: 10.1038/nrdp.2016.57
5. Héctor F. Escobar-Morreale, "Polycystic Ovary Syndrome," The Lancet Diabetes & Endocrinology, 2018. DOI: 10.1016/S2213-8587(18)30035-1
6. Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, 2001. DOI: 10.1214/aos/1013203451
7. Corinna Cortes and Vladimir Vapnik, "Support Vector Networks," Machine Learning, 1995. DOI: 10.1007/BF00994018
8. Tom Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, 2006. DOI: 10.1016/j.patrec.2005.10.010
9. Trevor Hastie et al., The Elements of Statistical Learning, 2009. DOI: 10.1007/978-0-387-84858-7
10. Gareth James et al., An Introduction to Statistical Learning, 2013. DOI: 10.1007/978-1-4614-7138-7
11. Pedro Domingos, "A Few Useful Things to Know About Machine Learning," Communications of the ACM, 2012. DOI: 10.1145/2347736.2347755
12. Kaur, H., et al., "Machine Learning-Based PCOS Detection," Journal of Biomedical Informatics, 2022. DOI: 10.1016/j.jbi.2022.104061
13. Sharma, M., et al., "AI-Based Healthcare Prediction Systems," IEEE Access, 2021. DOI: 10.1109/ACCESS.2021.3052100
14. Gupta, R., et al., "Predictive Analytics in Women Health," Applied Soft Computing, 2023. DOI: 10.1016/j.asoc.2023.109123
15. Singh, A., et al., "Ensemble Learning for Medical Diagnosis," Expert Systems with Applications, 2022. DOI: 10.1016/j.eswa.2022.117890