

XGBoost-Driven PCOS Prediction Using Multidimensional Clinical, Hormonal, and Lifestyle Features

Gourika J Sreenivas¹, Riya K P.², Yamini R.³

^{1,2}Students, ³Associate Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Chennai, India.

E-mail: ¹gs7893@srmist.edu.in, ²rb4841@srmist.edu.in, ³yaminir@srmist.edu.in

Abstract

Polycystic Ovary Syndrome (PCOS) is an endocrine-metabolic disease caused by various hormonal, clinical and lifestyle aspects, thus making its early prediction difficult. This research proposes an efficient predictive framework to classify PCOS by employing Machine Learning (ML) models that are trained with a dataset of 541 patient records with clinical, hormonal, physical, reproductive, lifestyle and other associated health characteristics. Four different supervised learning approaches such as Logistic Regression, Support Vector Machine, Random Forest, and Extreme Gradient Boosting (XGBoost) are compared based on the stratified five-fold cross validation process. The performance of the proposed model is evaluated in terms of accuracy, precision, recall, F1-score, and ROC-AUC. The highest performance is exhibited by the XGBoost model with an accuracy of 89.81%, precision of 93.10%, recall of 75.00%, and F1-Score of 83.08%, whereas the Random Forest attains highest ROC-AUC of 96.62%.

Keywords: Polycystic Ovary Syndrome (PCOS), Hormonal Feature Analysis, Lifestyle-Based Prediction, Extreme Gradient Boosting, Stratified Cross-Validation, Explainable Prediction, Probability-Based Risk Assessment.

1. Introduction

Polycystic Ovary Syndrome (PCOS) is an endocrine and metabolic disease commonly diagnosed in reproductive-age women. PCOS is characterized by the following symptoms: menstrual disorders, hormonal imbalance, overweight, insulin resistance, infertility, as well as

others. There is a significant variation in the manifestations of PCOS from one patient to another, which makes PCOS detection difficult. Information on LH, FSH, PRL, TSH, AMH hormone levels, as well as clinical and lifestyle characteristics of patients suffering from PCOS, can serve as an important source of information for disease diagnosis. Relationships among the above factors are complex and non-linear.

Traditional PCOS diagnosis usually depends on the methods of clinical examination, biochemical tests, and ultrasonographic observation. Despite the valuable information provided by such approaches, a simultaneous consideration of several heterogeneous features of a patient might be problematic. Machine Learning (ML) algorithms constitute a highly efficient pattern recognition in multidimensional healthcare data and can account for a wide variety of interdependencies between the variables. Some recent research works have already shown the applicability of ML for automated PCOS detection based on different feature sets [1], [2], [5].

The integration of various elements of patients can possibly provide a more complete picture of the features of PCOS. Specifically, the integration of hormonal markers and characteristics of patients in terms of their clinical and lifestyle can be useful in understanding both endocrine and metabolic features of the syndrome. It is crucial to consider different classification algorithms because they can behave differently depending on input characteristics [3], [4], [10].

Therefore, this research is focused on predicting PCOS using a dataset with features related to clinical, hormonal, and lifestyle aspects. Four ML classifiers, specifically LR, SVM, RF, and XGBoost, are analyzed for their performance using stratified fivefold cross-validation approach. Moreover, the current research aims to explore significant predictors and use the chosen classifier to predict PCOS probabilistically by calculating the percentage of PCOS probability.

2. Related Works

Recent researches have proven the effectiveness of the machine learning methods in predicting PCOS based on clinical, biochemical, hormonal, and lifestyle-based attributes. There is also a possibility to create an automated diagnosis system with the integration of supervised classification algorithms to classify PCOS and non-PCOS based on structured patient information [1], [2].

Feature selection and ensemble learning methods have become popular techniques for improving PCOS prediction in recent years. Feature selection methods have paid attention to choosing informative predictors and improving the interpretability of predictive models, whereas stacking techniques have concentrated on integrating several classifiers for better prediction performance [3], [4]. Clinical and EHR-based studies have additionally shown the potential of applying ML for PCOS patient-level prediction [5], [6].

Modelling challenges like class imbalance, hormonal feature selection, and high dimensional clinical data have also been tackled by recent researches [7], [8], [10]. Prediction using non-invasive methods like ultrasound with clinical variables has also shown the value of combining diagnostic modalities [9].

However, existing research works differ in terms of features considered, preprocessing techniques, and the selection of classifiers. While several works rely on hormonal, clinical, imaging, or electronic health records' parameters, there are only a few works that combine the effects of hormonal, clinical, metabolic, and lifestyle factors by systematically comparing different machine learning algorithms [2], [5], [8], [9]. Thus, in the current research, LR, SVM, RF, and XGBoost are applied to predict PCOS and the significance of features and model prediction probability are analyzed.

3. Proposed Methodology

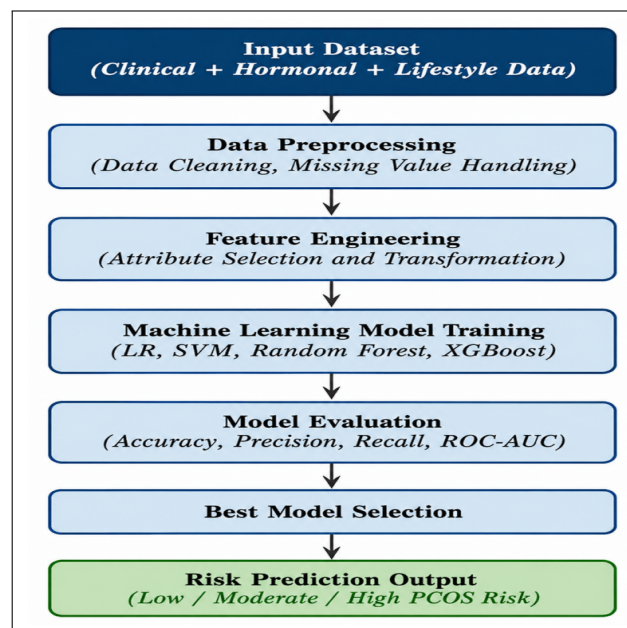


Figure 1. Integrated architecture of the machine-learning-based PCOS prediction framework

The proposed framework is meant for predicting PCOS through the combined analysis of clinical, hormonal, physical, reproductive, and lifestyle-related features. The approach is based on the following sequential machine learning pipeline that includes: dataset preparation; data pre-processing; feature engineering; classification; cross-validation; performance evaluation; feature importance; and probabilistic prediction. The general structure of the proposed framework is shown in Figure 1, where the patient information flow through the main computational steps is demonstrated starting from the initial dataset and finishing with the final PCOS prediction.

3.1 Dataset Description and Preprocessing

This research involves use of the public PCOS dataset [11] as the source of patient records. Despite the fact that the original source dataset has more records than those used in the study, only 541 records of patients have been selected to carry out the experiment. The aforementioned 541 records form the total dataset that needs to be analyzed in this experiment. The selected records have different patient attributes, ranging from clinical attributes, physical attributes, hormonal attributes, reproductive attributes, lifestyle attributes and ultrasonographic attributes.

The dataset includes several pertinent endocrine variables such as Anti-Müllerian Hormone (AMH), Luteinizing Hormone (LH), Follicle Stimulating Hormone (FSH), Thyroid-Stimulating Hormone (TSH), and Prolactin (PRL). Besides, physical and clinical variables like body mass index (BMI), menstruation features, follicular features, weight-based variables, and waist-to-hip ratio are included. Variables related to lifestyle are included to capture behavioral variables which could affect PCOS characteristics.

Prior to model training, the selected 541 instances undergo data preprocessing to achieve a uniform feature representation. Input features are divided into predictor features, whereas PCOS is isolated as the target feature. Feature engineering follows to preprocess and examine the available features for the discovery of meaningful predictor features. The feature set that is processed is finally supplied to the machine learning models.

3.2 Feature Engineering and Machine-Learning Model Development

After the data preprocessing phase, the 541 filtered patients' information is converted into the appropriate form of features for PCOS classification. The process of feature engineering involves the selection of relevant clinical, hormonal, physical, reproductive, and behavioral

features that would help in predicting PCOS. Special emphasis is put on those features related to hormonal and metabolic issues since they may have complicated relations with the condition under research.

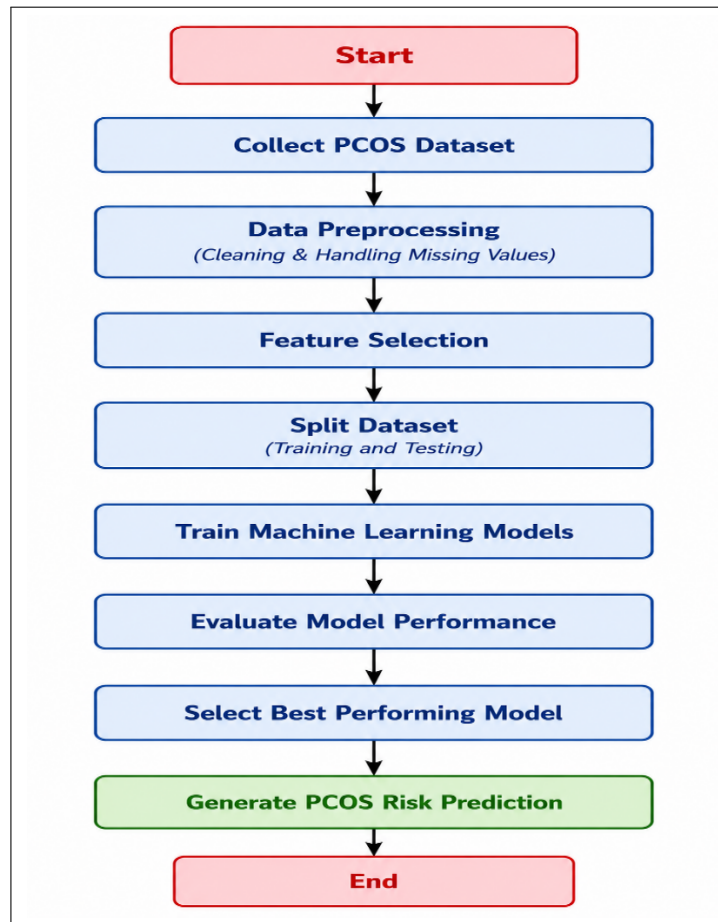


Figure 2. Sequential workflow of the machine-learning-based PCOS prediction process

As illustrated in Figure 2, the stage of feature processing is then followed by training and evaluating machine learning models. Four supervised classifiers have been used for the experiment, namely Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Each model provides a different approach to learn about the association between input features and PCOS status. In particular, LR creates a simple linear probabilistic classifier, SVM tries to find a good decision boundary separating the two classes, RF creates an ensemble of decision trees, and XGBoost uses boosting.

In order to get a reliable evaluation, the models are tested using stratified five-fold cross-validation as per the requirements of the research. The dataset is split into five folds in such a way that the ratio of the PCOS and non-PCOS classes remains the same. Four out of

these five folds are used for training in one round and one-fold is left for testing, and the process is repeated until every fold gets a chance to act as the test fold.

3.3 Model Evaluation and Selection

The four ML models will be validated by using stratified five-fold cross-validation, where the 541 patients are split up into five folds by retaining the same distribution of PCOS and non-PCOS patients. Four of the folds will be used to train the model and the fifth fold will be used as the validation data. This process is repeated five times such that each fold gets used for validation once.

Accuracy, Precision, Recall, F1-Score, and ROC-AUC are used for evaluating the model predictions. ROC-AUC is also used for measuring the ability of the models to discriminate between PCOS and non-PCOS samples. These performance metrics are measured for Logistic Regression, SVM, Random Forest, and XGBoost.

3.4 Feature Importance and Probability-Based PCOS Prediction

Following comparative assessment, the classifier that performs the best is chosen for the final prediction process. Feature importance assessment is carried out to determine which patient features play a significant role in the classification of PCOS. This method gives an extra insight into the model by assessing the relative contribution of hormonal, clinical, physical, and reproductive traits to the prediction process.

The analysis identifies AMH, LH, FSH/LH ratio, follicular characteristics, BMI, menstrual irregularity, and waist-hip ratio as important predictive features. These variables offer different pieces of information that relate to hormonal and metabolic aspects of PCOS.

Probability prediction of the PCOS category is made using the chosen model for each subject. The probability value is further expressed in percentage form in order to provide an understandable depiction of the estimated probability of PCOS:

$$\text{PCOS Probability (\%)} = P(\text{PCOS}) \times 100$$

These result percentage are classified into low, moderate, and high prediction levels based on the cut-off values set in the study. This probabilistic result generates an incremental prediction and not just a yes/no PCOS prediction.

4. Results and Discussion

The experimental evaluation was conducted in a Python-based machine-learning environment using 541 patient records from the PCOS dataset [11]. Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost were comparatively evaluated using stratified five-fold cross-validation. Model performance was assessed using the evaluation metrics defined in the proposed framework, and ROC analysis was used to examine class-discrimination capability.

4.1 Comparative Performance Analysis

The comparison results for all four classifiers have been tabulated in Table 1. XGBoost is able to give the highest values in terms of Accuracy (89.81%), Precision (93.10%), Recall (75.00%) and F1-score (83.08%). This makes XGBoost exhibit the best classification performance among all other classifiers. Random Forest gives Accuracy and F1-score of 87.96% and 79.37%, respectively, but gives the best ROC-AUC score of 96.6

The performance of the SVM model follows its own characteristic trend, with a 100% Precision score and a 2.78% Recall, giving an F1-score of 5.41%. This shows that the SVM model is able to predict the few cases which have been identified as positive, but cannot identify most of the cases having PCOS. On the other hand, the XGBoost algorithm gives a considerably higher recall rate along with high Precision score.

Table 1. Performance Comparison of Machine-Learning Models

ML Model	Model Accuracy	Model Precision	Model Recall	F1-Score	ROC-AUC
LR	0.8518	0.8846	0.6388	0.7419	0.9290
SVM	0.6759	1.0000	0.02777	0.0540	0.7110
Random Forest	0.8796	0.9259	0.6944	0.7936	0.9662
XGBoost	0.8981	0.9310	0.7500	0.8307	0.9656

4.2 Receiver Operating Characteristic Analysis

Discrimination ability of each classifier is analyzed through the Receiver Operating Characteristics (ROC) curves. Figure 3 shows the ROC curve generated for Logistic Regression. The ROC-AUC of 92.90% demonstrates that LR gives excellent separation of PCOS and Non-PCOS cases at various classification levels.

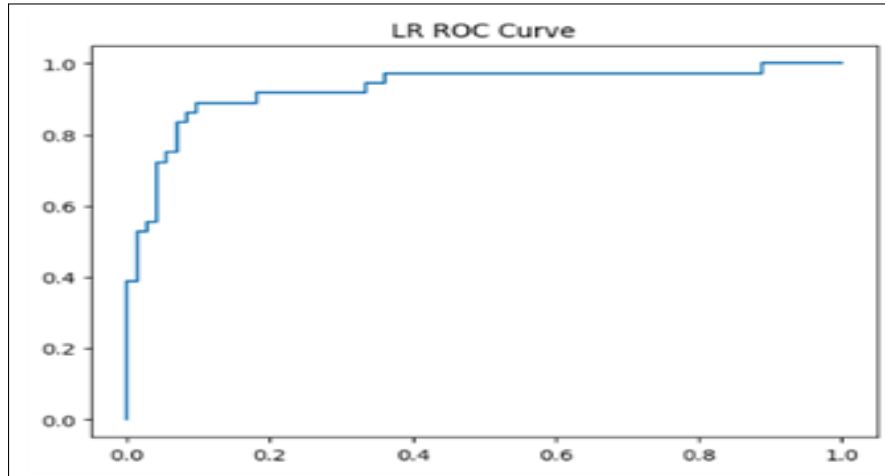


Figure 3. Receiver operating characteristic curve of logistic regression

Figure 4 provides the ROC analysis of Random Forest. RF manages to obtain a ROC-AUC score of 96.62%, which is the best discriminatory score among the models analyzed using this measure. This means that the ensemble method performs well in discriminating between classes at various thresholds.

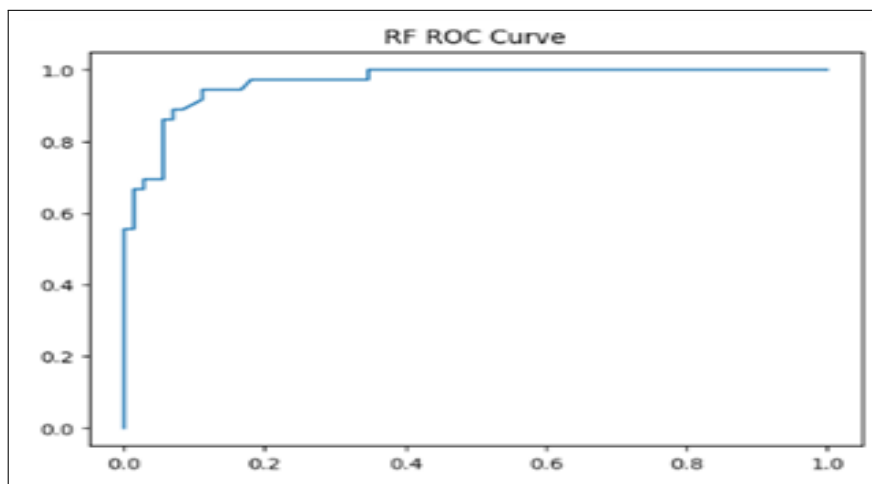


Figure 4. Receiver operating characteristic curve of random forest

Figure 5 below shows the ROC curve for the SVM classifier. The classifier has a ROC-AUC value of 71.10% which is significantly smaller compared to other classifiers. Reduced

discrimination power in this case is clearly illustrated by the extremely low Recall and F1-Score shown in Table 1.

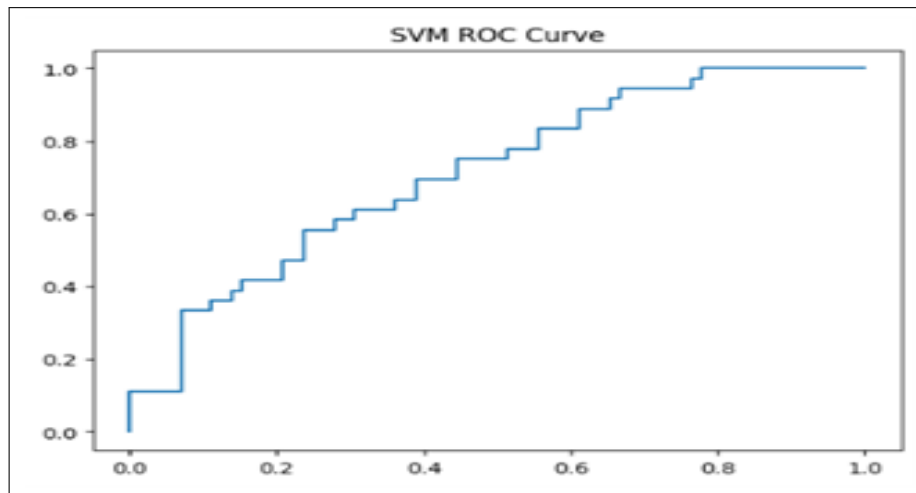


Figure 5. Receiver operating characteristic curve of support vector machine

Figure 6 shows the ROC curve for the XGBoost algorithm. The ROC-AUC score of the XGBoost algorithm is 96.57%, which denotes that XGBoost algorithm possesses high discriminative ability between the two classes. Despite having a ROC-AUC value slightly less than that of the RF algorithm, makes the XGBoost make it a superior classifier.

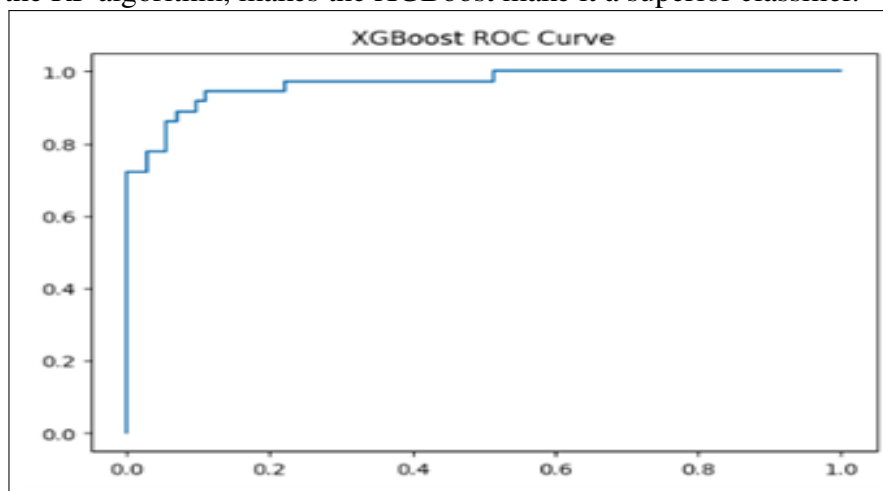


Figure 6. Receiver Operating Characteristic Curve of Extreme Gradient Boosting

4.3 Feature Importance Analysis

Individual contribution of each input feature to the prediction process is evaluated through the use of feature importance analysis.

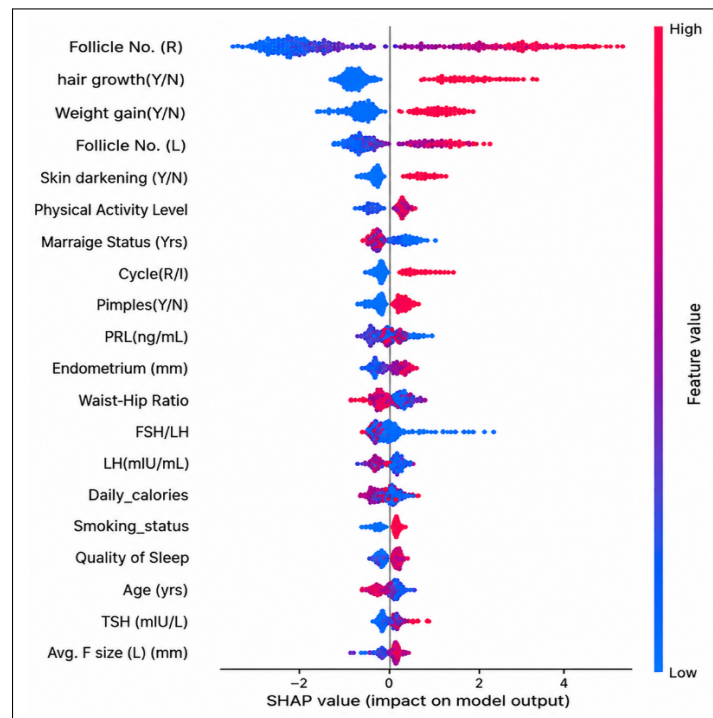


Figure 7. Feature importance of clinical and hormonal attributes for PCOS prediction

Figure 7 shows the importance of the selected features from the clinical, hormonal, physical, and lifestyle domains for the PCOS prediction process. The findings indicate that some of the most important features in relation to the model output include right-side follicle count, hair growth, weight gain, left-side follicle count, skin darkening, physical activity, marital status, menstruation, and pimples. Other important features include prolactin (PRL), endometrial features, waist-hip ratio, FSH/LH ratio, LH, daily caloric intake, smoking status, sleep pattern, age, TSH, and left-side average follicle size. The distribution of features reveals that the prediction of PCOS is affected by a range of factors related to reproductive, clinical, hormonal, metabolic, and lifestyle aspects of patients and not by any single aspect. It can be said that the significance of follicular and reproductive features in addition to hormones and lifestyle aspects confirms the multidimensionality of the problem. This outcome also confirms the aim of the research to incorporate different characteristics of patients to classify PCOS.

4.4 Probability-Based PCOS Prediction

Following model selection, the predictive framework generates a probability-based output for individual PCOS predictions. Figure 8 illustrates the variation in the model-estimated PCOS prediction with feature accumulation.

This behavior pattern shows that the calculated probability for PCOS does not rise linearly with more features added. This shows that the contribution from each particular variable to prediction relies on its overall presence in the system. This kind of behavior corresponds to the multi-dimensional aspect of the feature space, where different interactions from different aspects can affect the output classification.

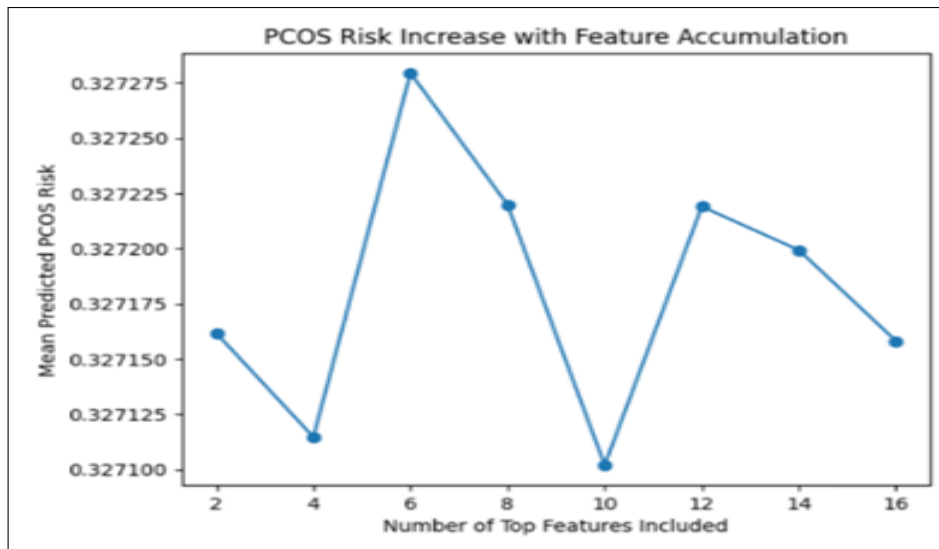


Figure 8. Variation in model-estimated PCOS probability with feature accumulation

The probability results from the model are expressed as percentage values and grouped under three different levels: low, medium, and high based on the cut-off points in the study. The findings in the analysis indicate that majority of the cases without PCOS have values less than 30%, while PCOS have values above 70%.

5. Conclusion and Future Work

This research study has proposed a PCOS prediction model using ML techniques by considering the clinical, hormonal, physical, reproductive and lifestyle-related parameters obtained from 541 patient cases. LR, SVM, RF and XGBoost were tested by using stratified five-fold cross-validation. The experiment results indicated that XGBoost has offered efficient predictive performance compared to other classifiers, whereas Random Forest was more competent in terms of discriminating ability between classes according to ROC-AUC. Feature importance analysis indicated the significance of hormonal, reproductive, physical and lifestyle parameters in predicting PCOS. The probability estimation approach provided an interpretable expression for the predicted probability of PCOS in addition to its classification. Future research

will concentrate on applying the proposed framework to independent larger datasets, probability calibration, including other clinically meaningful parameters and external clinical validation of the model.

References

- [1] Tiwari, Shamik, Lalit Kane, Deepika Koundal, Anurag Jain, Adi Alhudhaif, Kemal Polat, Atef Zaguia, Fayadh Alenezi, and Sara A. Althubiti. "SPOSDS: A Smart PCOSdiagnostic System using Machine Learning." *Expert Systems with Applications*, 2022, vol. 203: 117592.
- [2] Zigarelli, Angela, Ziyang Jia, and Hyunsun Lee. "Machine-Aided Self-Diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study." *JMIR Formative Research*, 2022, vol. 6, no. 3: e29967.
- [3] Elmannai, Hela, Nora El-Rashidy, Ibrahim Mashal, Manal Abdullah Alohal, Sara Farag, Shaker El-Sappagh, and Hager Saleh. "PCOSdetection Machine Learning Model based on Optimized Feature Selection and Explainable Artificial Intelligence." *Diagnostics*, 2023, vol. 13, no. 8: 1506.
- [4] Suha, Sayma Alam, and Muhammad Nazrul Islam. "Exploring the Dominant Features and Data-Driven Detection of PCOSthrough Modified Stacking Ensemble Machine Learning Technique." *Heliyon*, 2023, vol. 9, no. 3: e14518.
- [5] Zad, Zahra, Victoria S. Jiang, Amber T. Wolf, Taiyao Wang, J. Jojo Cheng, Ioannis Ch Paschalidis, and Shruthi Mahalingaiah. "Predicting PCOSwith Machine Learning Algorithms from Electronic Health Records." *Frontiers in Endocrinology*, 2024, vol. 15: 1298628.
- [6] Rahman, Md Mahbubur, Ashikul Islam, Forhadul Islam, Mashruba Zaman, Md Rafiul Islam, Md Shahriar Alam Sakib, and Hafiz Md Hasan Babu. "Empowering Early Detection: A Web-based Machine Learning Approach for PCOS Prediction." *Informatics in Medicine Unlocked*, 2024, vol. 47: 101500.
- [7] Ahmad, Rumman, Lamees A. Maghrabi, Ishfaq Ahmad Khaja, Louai A. Maghrabi, and Musheer Ahmad. "SMOTE-based Automated PCOS Prediction using Lightweight Deep Learning Models." *Diagnostics*, 2024, vol. 14, no. 19: 2225.

- [8] Tong, Cheng, Yue Wu, Zhenchao Zhuang, and Ying Yu. "A Diagnostic Model for PCOSbased on Machine Learning." *Scientific Reports*, 2025, vol. 15, no. 1: 9821.
- [9] Agirsoy, Mehtap, and Matthew A. Oehlschlaeger. "A Machine Learning Approach for Non-Invasive PCOS Diagnosis from Ultrasound and Clinical Features." *Scientific Reports*, 2025, vol. 15, no. 1: 33638.
- [10] Acharya, Sachin, Satyanarayana Poojari, and Asha Kamath. "Prediction of PCOSusing Machine Learning Models: Addressing Class Imbalance and High Dimensionality." *Journal of Education and Health Promotion*, 2026, vol. 15: 174.
- [11] Kottarathil, Prasoon. PCOS(PCOS). Kaggle Dataset, 2020. Available:
<https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>