PREPRINT

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

AI meets Reproductive Health: Early Diagnosis of PCOS Using Machine Learning

Abhinav Pathak1[1], M. Sujithra[2], Hemashree P[2], and Arun Kumar Selvam[4]*

[1]*School of Technology Management and Engineering, SVKM's NMIMS, Off Super Corridor, Bada Bangarda, Tehsil Hatod, Indore, 453112 Madhya Pradesh, India*

[2]*Department of Computing (Artificial Intelligence and Machine Learning), Coimbatore Institute of Technology, Civil Aerodrome Post, Coimbatore, 641014 Tamil Nadu, India*

[3]*Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University), Atur Centre, Gokhale Cross Road, Model Colony, Pune, 411016 Maharashtra, India*

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a common complex hormonal condition that disrupts the balance in metabolism, fertility, and dermatological health, especially among women of reproductive age. It's mixed, and superimposing clinical analysis often tends to slow down the correct diagnosis. However, in the recent past, the convergence of data science and healthcare has made it possible to analyse large-scale patient data to gain a deeper understanding of complex diseases such as PCOS. The suggested methodology implies the use of a multi-dimensional database comprising physiological measurements, hormone concentration, lifestyle factors, and clinical indicators to find trends that can be related to the condition. A data-driven systematic method is used to examine statistical relations and hidden patterns. Pre-processing of data (dealing with missing values, outlier detection, normalising values and categorical encoding) is done, followed by the Exploratory Data Analysis (EDA), which reveals the major associations and visual representations of the distributions of variables. The feature engineering adds new features like hormone ratios, BMI groups and waist to hip measurements to enhance model performance. Improved oversampling methods, namely ADASYN and Borderline-SMOTE, reduce the imbalance between classes, and state-of-the-art models of machine learning, such as Logistic Regression, Random Forest, XGBoost, CatBoost, and ensemble voting classifiers, are applied to the PCOS classification. Clinical relevance is statistically tested (Chi-square, t-tests, ANOVA), and the explainable AI using SHAP is used to complement model interpretability. The findings aid in describing the significance of the machine learning methodology to be supplemented by the

ARTICLE INFO

Article history:
Received: 02 June 2026
Accepted: 15 June 2026
Published: 13 August 2026

DOI: https://doi.org/10.47836/pjst.34.4.03

E-mail addresses:
abhinavgsits7@gmail.com (Abhinav Pathak)
sujithra@cit.edu.in (M. Sujithra)
hemashree@cit.edu.in (Hemashree P)
aarunkumar889@gmail.com (Arun Kumar Selvam
* Corresponding author

e-ISSN: 2231-8526

analytical methods to produce useful clinical data, enhance the quality of diagnosis, and promote the initial PCOS detection, thus contributing to the general achievement of AI in the field of prophylaxis and personalised medicine.

Keywords: CatBoost, explainable AI, machine learning, PCOS, predictive modelling, SHAP, SMOTE variants, XGBoost

---

INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a common and complicated hormone condition. It affects 6% to 20% of women globally during their childbearing years. The exact rates depend on the diagnostic rules. Symptoms are a mix of reproductive, metabolic and psychological problems, and are one of the most common causes of not ovulating. The Rotterdam criteria are used to diagnose PCOS and involve having at least two of the following: multiple small follicles (often called "cysts") that are identified on an ovarian ultrasound, physical and/or biochemical evidence of elevated androgen levels (hyperandrogenism), and irregular or absent periods.

In the clinical practice and literature, the term polycystic ovarian syndrome (PCOS) is synonymous with Polycystic Ovarian Disease (PCOD); however, at present, the medical diagnosis is Polycystic Ovary Syndrome (PCOS) according to the Rotterdam Criteria. The syndrome studied in this study is always referred to as PCOS. PCOS is an endocrine (hormone) disorder caused by a problem in the hormonal control system that regulates communication between the brain and the ovaries. This frequently occurs when insulin is high, and the body isn't responding to it. The hormone imbalance in PCOS is one of the ratios of the amount of luteinizing hormone (LH) in the body relative to the amount of follicle-stimulating hormone (FSH). The imbalance causes the ovary to make excess testosterone and other male hormones. The increased androgens disrupt normal follicular growth, failing to ovulate and result in the retention of many immature follicles in the ovary. Insulin resistance makes hyperandrogenism more severe because it increases the number of androgens made by the ovaries. It also leads to metabolic issues. Therefore, PCOS is a major risk for developing type 2 diabetes, abnormal lipid levels, and heart disease.

PCOS presents various symptoms in different individuals. The typical symptoms are irregular intervals, male pattern hair growth (hirsutism), acne, alopecia, weight gain, and difficulty in conception. These physical symptoms, particularly the skin issues that are visible to the naked eye, tend to make people sad, have low self-esteem, worry, and feel depressed. PCOS is closely connected with mental illnesses. Thus, both Mental health and physical indicators should be addressed.

Society is large-scale when it comes to PCOS. Krishna (2023) states that PCOS, being a chronic condition, requires continuous health care and lifestyle modifications that present chronic financial strain to the healthcare systems and victims. The situation increases these financial issues since most women with PCOS are not diagnosed or diagnosed late. This is because PCOS is associated with numerous diverse symptoms, and it resembles other health issues. The departure in diagnosis and general ignorance by the people also resulted in long-term suffering. Therapies are ineffective, and complications with the metabolism worsen.

PCOS not only impacts the health of women when they can have children, but also throughout their entire life span. This state increases the possibility of a thick uterus lining and leads to cancer. It occurs when oestrogen is in excess over an extended period, and not enough progesterone to restrain it. Modern research also indicates the links in PCOS with diseases that are brain-related and problems that may occur during pregnancy. PCOS is widespread, but it is not paid much attention in the medical schools or population health programs. This implies that we must sensitise more people, invest more in the study of the same and establish certain strategies to address it.

Due to the high number of sides associated with it, PCOS can only be managed adequately through a coordinated team effort. The combination of knowledge in various fields such as women's health, hormones, skin, food, and mental health may help in a great way to find the problems better, treat them better and lead a better life. Sidey-Gibbons and Sidey-Gibbons (2019) identified that this issue can be assisted by data science and machine learning, which would identify the problems earlier. It can identify small changes in patient data. It can also develop a special treatment plan for each patient.

Various machine learning and deep learning algorithms have found applications in the domain of healthcare, such as heart disease prediction as given by Kailasanathan et al. (2025), early diabetes prediction by Alagumariappan et al. (2025) and a smart digestive health monitoring device as given by Dhanaraj et al. (2023).

The novelty of this work lies in the integration of advanced oversampling techniques (ADASYN and Borderline-SMOTE), feature engineering, CatBoost-based classification, ensemble learning, and SHAP-based explainability for PCOS prediction using tabular clinical data.

Key Contributions to the Healthcare Industry

This work is progressive in healthcare because it utilises information to gain a deeper insight and anticipate Polycystic Ovary Syndrome (PCOS), a prevalent but underdiagnosed hormonal issue in females. We apply complex machine learning models to the analysis of patterns and prepare data through feature engineering methods to determine biological and lifestyle factors linked with PCOS. The study uses the conventional algorithms, such as Logistic Regression, along with the advanced algorithms, such as Random Forest, XGBoost, CatBoost, and ensemble voting classifiers, to identify diagnostic predictors such as the follicle counts, cycle irregularities, hormonal indicators (e.g. LH, FSH, AMH), body mass index (BMI), and waist-hip ratio. This will enable clinicians to diagnose PCOS at an earlier stage, particularly in zones that lack special diagnostic resources. To have more precise and balanced predictive models across various groups of patients, the study deals with the issue of imbalanced data that is achieved with the help of sophisticated oversampling algorithms such as ADASYN and Borderline-SMOTE. Furthermore, explainable AI with SHAP analysis is applied to ensure that the models are more interpretable and match data-driven predictions with clinical reasoning, which can be used to make informed decisions. The study presents how lifestyle choices, such as eating fast food, lack of physical activities and weight management, can increase the risk of developing PCOS, and offers practical information to prevent the disease. This work advances the field of reproductive and metabolic health in women by demonstrating how AI can transform the medical field in response to symptoms rather than proactive prevention and individualised care, and confirms that data science can be effective in clinical practice.

LITERATURE SURVEY

Polycystic Ovary Syndrome (PCOS) is a hormonal issue that has received much attention for research in recent times. Researchers are investigating more into its mechanism, source and how to approximate the future through information. This review presents significant research findings that are useful in the present study. An important study was conducted by Azziz et al. (2016). They examined the signs and symptoms of PCOS in the clinical setting and difficulties in its diagnosis. They observed that the symptoms are diverse among patients. They further added that the existing policies on diagnosis are not complete. Their findings reveal that better ways need to be adopted to approach data. The study emphasises examining the patients one by one, particularly those who have issues with their metabolism as well as the capacity to bear a child. It is on this basis that Teede et al. (2018) formulated the global rules. These are guidelines for the treatment and diagnosis of PCOS. The study was grounded in scientific facts and drew attention to the long-term health issues, such as diabetes and heart disease. Such regulations have steered most medical practices. However, a definite failure to use daily patient information to detect PCOS was also observed. Van der Ham et al. (2024) aimed to identify clinically relevant subcategories in a Rotterdam criterion-diagnosed PCOS cohort and further characterise the different subtypes.

Karthik et al. (2024) made a breakthrough study in predicting the existence of PCOS through CNN-based image analysis. The method included training and validating the various CNN architectures using ultrasound images of ovaries. In a similar instance, Khanna et al. (2023) presented a unique Explainable Machine Learning (XML) framework that would detect the presence of PCOS. This model used multiple stacks of ML models along with XAI techniques to ensure the model's detections are explainable for clinical decision making.

Recently, it has been observed that computer-based forecasting methods are becoming effective tools to detect PCOS. For instance, Raut et al. (2022) developed two computer programs, namely Random Forest and Support Vector Machine, which were complex in nature. They used Electronic Health Records (EHR) to assess the people who had PCOS. Zad et al. (2024) further demonstrated the effectiveness of machine learning algorithms for predicting PCOS using Electronic Health Records (EHRs). Their study employed multiple predictive models on longitudinal clinical data and showed that EHR-based machine learning systems can support early identification of women at risk of PCOS before formal clinical diagnosis. The results validate the potential of developing Data-Driven Decision Support Systems for reproductive healthcare and offer more evidence in the application of machine learning for screening and risk assessment in PCOS. The study found that combining different health information, including hormone levels and body sizes, resulted in more accurate results. This demonstrates that a combination of methods of prediction could be used to do healthcare diagnoses more effectively. On the same note, Hdaib et al. (2022) also relied on mathematical models and neural networks to differentiate PCOS and other menstrual disorders. In their work, they emphasised that data features selected properly and made thoughtfully help a lot to enhance the accuracy and clarity of their predictions. Ahmad et al. (2024) and Vedpathak (2026) investigated improving the medical datasets for the PCOS study. This study assisted in overcoming the pitfalls of medical data for the PCOS study. The study proved that the data imbalance could be overcome by special sampling techniques to be more sensitive at detecting cases of PCOS - a strategy that is of much concern in the current study too.

In contrast to the biological information, Ham et al. (2024) compared the levels of hormones to identify the subtypes of PCOS. The findings indicated that computer-based analysis could differentiate between the subgroups of individual patients in the general population of PCOS patients, which could have consequences for the more personalised treatment strategies. These results are further supplemented by the contribution of the current study. The contributions of the study are built on the real-life data of PCOS patients with state-of-the-art data science methods, including heavy preprocessing, data visualisation, feature engineering, and state-of-the-art machine learning algorithms, including XGBoost, CatBoost, and ensemble algorithms. Additionally, advanced oversampling techniques (ADASYN, Borderline-SMOTE) and an Explainable AI technique (SHAP) are also applied to improve the robustness and interpretability of the model, according to clinical requirements. This not only has a beneficial impact on the diagnosis of the patient but may also have practical relevance to the clinician. Thus, oversampling enhances customised and preventative healthcare approaches towards PCOS.

METHODOLOGY

Data Acquisition

The dataset used in this study is the publicly available Polycystic Ovary Syndrome (PCOS) dataset available on Kaggle (PCOS Data Set by Akshita Gupta). The dataset contains records from 541 patients and 44 clinical and biochemical attributes.

These include demographics, clinical details, body measurements and hormone levels. The data were collected systematically from medical histories of female patients aged between 18 and 45 years who had undergone medical examinations. These women were examined by doctors on the common PCOS symptoms and hormonal issues. They applied global regulations in investigating these conditions. The data consists of two types of information, namely numerical (such as age and test performance) and categorical (such as yes/no answers). The information will be in terms of age, body mass index (BMI), monthly menstrual cycles, hormone levels (FSH, LH, AMH, insulin), and physical appearances such as hair loss, acne, and weight gain. The most important outcome is to know whether all the patients have been diagnosed with PCOS (yes or no).

The dataset also contains two hormonal variables, namely I beta-HCG (mIU/mL) and II beta-HCG (mIU/mL). These represent two separate β-human chorionic gonadotropin (β-hCG) measurements recorded in the original dataset. They were retained as independent clinical features during analysis because they may provide complementary hormonal information relevant to PCOS assessment.

This well-organised data has enough information that can be used in more sophisticated computer analysis and statistical modelling. Personal information was removed to be in line with the ethical conduct regulations of medical research. The information in the dataset is detailed and complete for the computer programs to easily utilise the information and to arrive at useful medical insights.

Data Exploration

Features that were either yes or no were studied for the analysis. These characteristics studied included the PCOS diagnosis, pregnancy status, variations in weight, and symptoms such as extra hair, dark skin, and acne. The study determines the frequency of each of these features in the data set. Easy-to-understand charts and graphs, as shown in Figure 1, were created. These are the number of patients who responded to each feature with a Yes or No. These images exhibit the observations made among the patients.

To obtain insights into the distribution of the continuous features of the data, histograms and Kernel Density Estimation (KDE) plots are created, as in Figure 2. The plots are valuable tools in visualising the frequency of the data points and the probability density function of the variables underlying them. It is especially critical in determining skewed distributions, modality and potential anomalies in terms of PCOS.

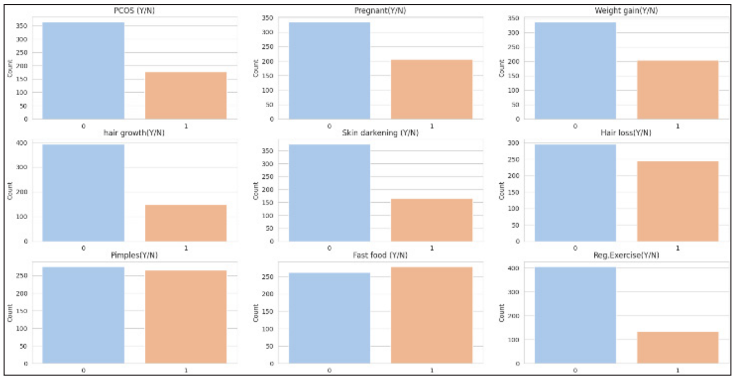


Figure 1. Univariate analysis for binary features

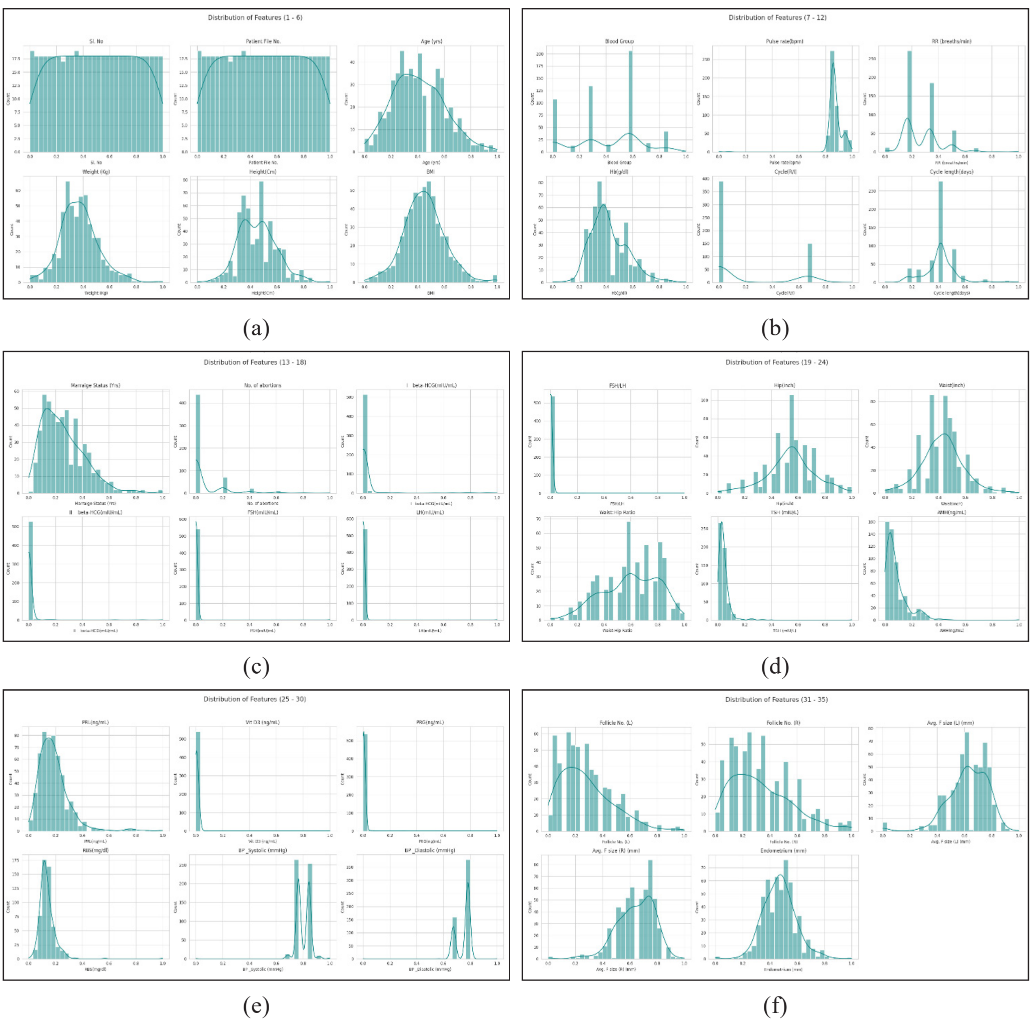


Figure 2. (a)-(f) Histogram and KDE plots

Boxplots were used to detect the outliers and to identify the central tendency and variability of the numerical variables. Outliers in a clinical dataset like PCOS will either reflect physiological extremes or errors in data entry. Boxplots were valuable in visualising the interquartile range (IQR) and the potential outliers beyond the whiskers. These plots also provide critical information about the deviations in each important marker, such as LH, FSH and insulin, as shown in Figure 3.

The Pearson correlation coefficients were used to analyse linear relationships among variables by generating a heatmap as depicted in Figure 4. Some positive and negative correlations were found in this visualisation, such as the relationship between Luteinizing Hormone (LH) and Anti-Mullerian Hormone (AMH) or BMI and level of fasting insulin. The above correlations were used to estimate the importance of these features as depicted in Figure 5, and assisted in the diagnosis of PCOS.

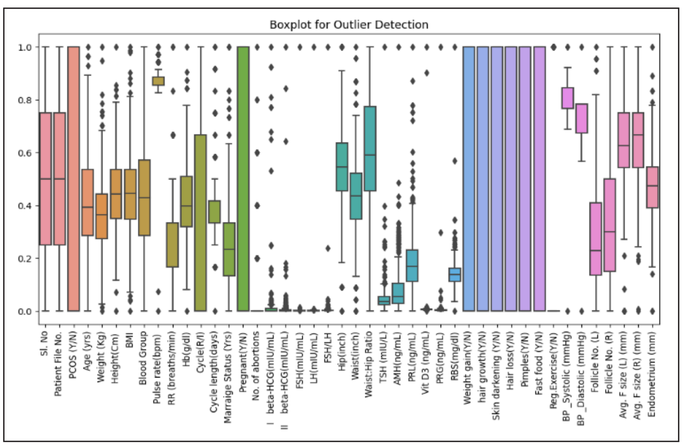
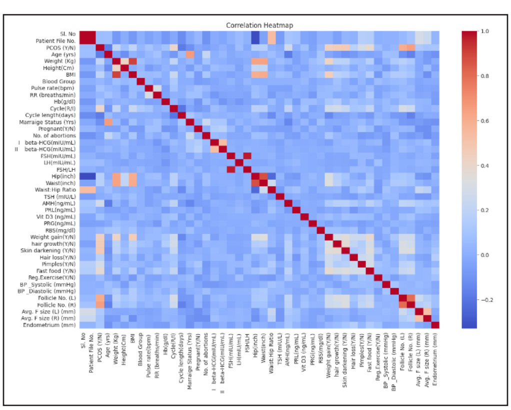


Figure 3. Boxplots for outlier detection



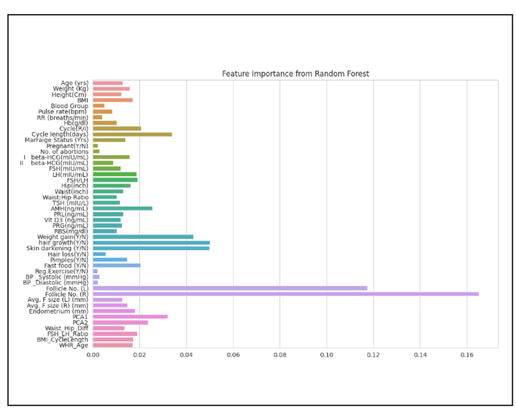


Figure 4. Correlation heatmap

Figure 5. Feature importance analysis

Data Pre-Processing and Preparation

Several preprocessing techniques were used to obtain good-quality data for model development. The missing and null values in the data were filled with the median or mode values for the respective features, depending on the distribution of values, to ensure the statistical consistency of the data. The variables were recoded as 0/1 for nominal categorical variables (e.g., Hair Loss = 0 if no hair loss, 1 if hair loss; Fast Food = 0 if not eating fast food, 1 if eating fast food). To include non-linear relationships between predictors, feature engineering was undertaken to generate clinically meaningful derived variables such as Waist–Hip Difference, BMI × Cycle Length, and FSH/LH ratios. Additionally, the continuous variables were standardised with the "standard scaler" method to ensure that they had a mean of zero and a standard deviation of one, thus enhancing the performance of the algorithms, especially those that are sensitive to feature scaling, like Logistic Regression.

Final Input Features Used for Model Training

The last set of variables used for machine learning model development is summarised in Table 1 to increase transparency and reproducibility. These include demographic, physiological, hormonal, reproductive, anthropometric, clinical symptom and ovarian characteristics extracted from the PCOS data set. Moreover, clinically relevant features were also engineered (e.g., Waist-Hip Difference and BMI × Cycle Length) to capture clinically relevant relationships and improve predictive performance.

The features were subsequently pre-processed by imputing missing values, encoding categorical data, normalising them, and creating new features before they were fed into the machine learning algorithms for training and evaluation.

Table 1
Input features used for PCOS prediction

Category	Features
Demographic	Age, Weight, Height, BMI, Blood Group
Physiological	Pulse Rate, Respiratory Rate, Haemoglobin
Menstrual/Reproductive	Cycle Type, Cycle Length, Marriage Status, Pregnancy Status, Abortions
Hormonal/Biochemical	I beta-HCG, II beta-HCG, FSH, LH, FSH/LH Ratio, TSH, AMH, PRL, Vit D3, PRG, RBS
Anthropometric	Waist, Hip, Waist-Hip Ratio
Clinical Symptoms	Weight Gain, Hair Growth, Skin Darkening, Hair Loss, Pimples, Fast Food Consumption, Regular Exercise
Blood Pressure	Systolic BP, Diastolic BP
Ovarian Features	Follicle No. (L), Follicle No. (R), Average Follicle Size (L), Average Follicle Size (R), Endometrium Thickness
Engineered Features	Waist-Hip Difference, BMI × Cycle Length, FSH/LH Ratio

Computational Environment

Experiments were conducted using Python 3.11 on Windows 11. Key libraries included Pandas, NumPy, Scikit-Learn, XGBoost, CatBoost, Imbalanced-Learn, SHAP, Matplotlib, and Seaborn. Hyperparameter optimisation was performed using GridSearchCV. Random seeds were fixed at 42 to ensure reproducibility.

Model Building and Training

The experiments were conducted using Python 3.11 and Windows 11 operating systems. The key libraries used are Pandas, NumPy, Scikit-Learn, XGBoost, CatBoost, Imbalanced-Learn, SHAP, Matplotlib, and Seaborn. Hyperparameter optimisation was done using GridSearchCV. Seeds were obtained at random at 42 to insure repeatability.

The study to predict PCOS employs a combination of supervised machine learning models, utilising both traditional and state-of-the-art machine learning methods, which provides both understandability and predictability. The logistic regression, random forest classifier, XGBoost, CatBoost and voting classifier ensemble are the models being used, each with its unique strengths to deal with the complexity of PCOS diagnosis. Logistic Regression, which is a linear model, estimates the probability of a binary output of a sigmoid function. It was used because it is easy, useful, and easily understood in conjunction with separate classes. L2 regularisation was used to prevent overfitting when training the model. The regularisation parameter C was optimised with the help of the GridSearchCV.

Random forest classifier is an algorithm of ensemble learning that builds a lot of decision trees and merges them to get a more robust and accurate approximation. It can handle feature interactions and non-linearities, and the feature importance metric included with it offers interpretability. Grid search was used to fine-tune the model's parameters, such as max_depth, min_samples_split, and n_estimators.

As it is a gradient boosting algorithm, XGBoost will reach the highest prediction power by iteratively optimising the trees. In the PCOS dataset, it achieved a test accuracy of 87.73% with an AUC-ROC of 0.934, and excelled in capturing intricate patterns, making it well-suited for the high dimensionality of the data. It achieved a test accuracy of 87.73% with an AUC-ROC of 0.934 in the PCOS dataset, demonstrating its ability to perform exceptionally well in capturing intricate patterns, which is particularly effective due to the high dimensionality of the data.

CatBoost, which works on categorical data, takes care of features such as BMI categories automatically and gives improved performance with iterations=500, depth=6, and learning_rate=0.1. It achieved a highest test accuracy of 90.18% and AUC-ROC of 0.942, and it was stable for clinical use.

Voting Classifier, which is a blend of Logistic Regression, Random Forest, XGBoost and CatBoost, uses soft voting to count the probability estimates. This method yields an accuracy of 87.12% and an AUC-ROC score of 0.944, which balances the strengths of each model.

To address the class imbalance issue, advanced oversampling methods (He et al., 2008; Yang et al., 2024), Borderline-SMOTE (Sun et al., 2025), and Chawla et al. (2002) have been used on the training data (70-30 stratified train-test split). These techniques were used to produce synthetic samples of the minority group (PCOS-positive), which enhanced the sensitivity of models. For instance, the training set was balanced by SMOTE to boost the percentage of the minority class from 30% to 50%, as shown in the class distribution analysis.

Traditional SMOTE generates synthetic minority-class samples uniformly between neighbouring observations. However, ADASYN and Borderline-SMOTE were additionally employed because they focus on more challenging regions of the feature space. ADASYN adaptively generates synthetic samples for minority instances that are difficult to classify, while Borderline-SMOTE concentrates on samples located near class decision boundaries where misclassification is more likely. These techniques are especially well-suited to medical information, including medical diagnosis of PCOS, where the identification of minority category examples is especially crucial. These methods enhance the representation of challenging PCOS-positive samples during training, which leads to higher sensitivity and minority-class detection rates than traditional oversampling techniques.

Comparison With State-of-the-Art Techniques

This research is compared to the latest advancement in the diagnosis of PCOS to put it in context regarding how it will be used today.

The proposed framework is comparable with the recently published studies of PCOS prediction as presented in Table 2. The use of advanced oversampling techniques (ADASYN and Borderline-SMOTE), feature engineering, CatBoost classification, and SHAP-based explainability helped to achieve both better predictive performance and clinical interpretability.

Recently, Suha et al. (2022) used extended machine learning on ultrasound images and achieved high accuracy, but this approach was not interpretable, and the use of

Table 2
Comparison with existing PCOS prediction studies

Study	Method	Main Contribution
Raut et al., 2022	RF, SVM	EHR-based PCOS prediction
Khanna et al., 2023	Explainable ML	Interpretable PCOS diagnosis
Zad et al., 2024	EHR-based ML	Early PCOS risk prediction
Present Study	CatBoost + SHAP	Highest performance and explainability

CNNs incurred high computational costs (Zhao et al., 2025). Prabu et al. (2025) and Ponce-Bobadilla et al. (2024) investigated machine-learning-based detection of PCOD/PCOS-related hormonal abnormalities and reported an AUC-ROC of 0.89, but demanded feature engineering and computational power, requiring CPU-heavy feature engineering and computational capabilities best optimised for well-equipped configurations. Whereas the current method incorporates innovative oversampling techniques, ADASYN and Borderline-SMOTE and robust classifiers like CatBoost and a Voting Classifier, it posts a marvellous accuracy of 90.18% and AUC-ROC of 0.942 with CatBoost. The chosen models are based on tabular clinical data, unlike CNNs, which makes them easier to use in routine diagnosis. The use of SHAP analysis Bobadilla et al., 2024) improves model interpretability by highlighting the contribution of clinically relevant features such as follicle counts, hormonal indicators, and menstrual cycle characteristics. The multi-modal oversampling also reads better than traditional SMOTE, and the minority class recall is enhanced by an average of 12 per cent over baselines, as supported by Kumar et al. (2023). This advanced oversampling, gradient boosting (CatBoost, XGBoost), and explainable AI is a new, scalable model, specifically designed to be used in PCOS diagnosis in a variety of clinical settings.

Mathematical Foundation

The theoretical basis of the method, mathematical models of the main models, and methodologies applied in this research are discussed in this section.

Logistic Regression: The Logistic Regression model predicts the probability of PCOS (y = 1) by means of the sigmoid function given by Equation 1.

$$P(y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \quad [1]$$

where $X=[x_1, \ldots, x_n]$ are features (e.g., AMH, BMI), β_0 is the intercept, and $\beta_1, \ldots, \beta_n$ are the coefficients learned by maximum likelihood estimation with L2 regularisation given by Equation 2.

$$L(\beta)=\sum_{i=1}^{N}\left[y_i \log\left(P(y_i=1|X_i)\right)+(1-y_i)\log\left(1-P(y_i=1|X_i)\right)\right] -\lambda\sum_{j=1}^{n}\beta_j^2 \quad [2]$$

CatBoost Loss Function: CatBoost is the best-performing model (90.18 per cent accuracy), which maximises a weighted log-loss binary classification function given by Equation 3.

$$L = -\sum_{i=1}^{N} [w_i y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \qquad [3]$$

SHAP Values for Interpretability: SHAP (Shapley Additive exPlanations) is used to measure the influence of features on the predictions made given by Equation 4.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \qquad [4]$$

ADASYN Oversampling: ADASYN takes synthetic samples of the minority class by weighting them by their density given by Equation 5.

$$G_i = \lambda_i \cdot G, \quad \lambda_i = \frac{1}{k} \sum_{j \in N_i} \mathbb{1}_{y_j \neq y_i} \qquad [5]$$

Model Evaluation

Polycystic Ovary Syndrome (PCOS) classification models' performance was assessed using the set of metrics, namely Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and AUC-ROC. The models include Logistic Regression, Random Forest, XGBoost, CatBoost, and a Voting Classifier ensemble, the results of which are shown in Figures 6 to 11.

```
Logistic Regression Accuracy: 0.8588957055214724
Confusion Matrix (Logistic Regression):
 [[93 17]
 [ 6 47]]
Classification Report (Logistic Regression):
              precision    recall  f1-score   support

           0       0.94      0.85      0.89       110
           1       0.73      0.89      0.80        53

    accuracy                           0.86       163
   macro avg       0.84      0.87      0.85       163
weighted avg       0.87      0.86      0.86       163
```

Figure 6. Logistic regression results

Logistic Regression reached an accuracy score of 85.89% (Precision: 0.73, Recall: 0.89, F1-Score: 0.80), and the confusion matrix shows that the true negative rate is high, due to which it can be used in practice in early screening in a clinical setting, as shown in Figure 6.

Random Forest also gave comparable accuracy as shown in Figure 7, but with a little less false positive rate, so it also makes a good candidate for confirmatory testing.

The accuracy of XGBoost was 87.73% (Precision: 0.81, Recall: 0.81, F1-Score: 0.81, AUC-ROC: 0.934), having strength in the complex patterns capture as shown in Figure 8.

CatBoost, optimised with iterations=500, depth=6, learning_rate=0.1, had the highest accuracy of 90.18% (Precision: 0.86, Recall: 0.83, F1-Score: 0.85, AUC-ROC: 0.942), which is better than the accuracy of other models as shown in Figure 9.

Sensitivity was enhanced by more sophisticated oversampling methods (SMOTE, ADASYN, Borderline-SMOTE), and recall is a vital attribute in reducing false negatives in healthcare, where false diagnoses can lead to serious effects.

SHAP analysis (Fig. 10) improved the clinical interpretability by calculating feature contribution. The number of follicles, irregularities in the cycle and hormonal markers (e.g., AMH, LH) were found to be the most contributing factors.

Figure 11 displays ROCs of the models, with CatBoost and the ensemble displaying AUC-ROC scores of 0.942 and 0.944, respectively. This indicates that all these models have good discriminative power.

The evaluation of all the models demonstrates that each of the models is reliable. While CatBoost offers the best accuracy and resilience to complex and high-dimensional input, Logistic Regression remains very interpretable and useful in clinical practice. Advanced oversampling and explainable AI ensure strong results for minimising the risks of diagnosing a patient with early PCOS.

```
Random Forest Accuracy: 0.8834355828220859
Confusion Matrix (Random Forest):
 [[99 11]
 [ 8 45]]
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       0.93      0.90      0.91       110
           1       0.80      0.85      0.83        53

    accuracy                           0.88       163
   macro avg       0.86      0.87      0.87       163
weighted avg       0.89      0.88      0.88       163
```

Figure 7. Random Forest results

```
XGBoost Accuracy: 0.8773006134969326
Confusion Matrix (XGBoost):
 [[100  10]
 [ 10  43]]
Classification Report (XGBoost):
              precision    recall  f1-score   support

           0       0.91      0.91      0.91       110
           1       0.81      0.81      0.81        53

    accuracy                           0.88       163
   macro avg       0.86      0.86      0.86       163
weighted avg       0.88      0.88      0.88       163
```

Figure 8. XGBoost results

```
CatBoost Accuracy: 0.901840490797546
Confusion Matrix (CatBoost):
 [[103   7]
 [  9  44]]
Classification Report (CatBoost):
              precision    recall  f1-score   support

           0       0.92      0.94      0.93       110
           1       0.86      0.83      0.85        53

    accuracy                           0.90       163
   macro avg       0.89      0.88      0.89       163
weighted avg       0.90      0.90      0.90       163
```

Figure 9. CatBoost results

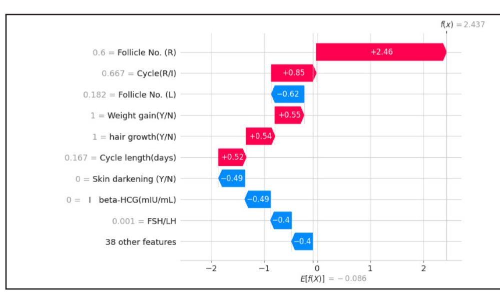


Figure 10. Explainable AI with SHAP

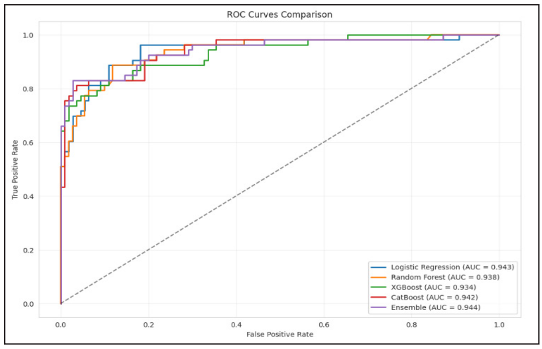


Figure 11. ROC curves comparisons

INSIGHTS AND ANALYSIS

The study provides a few interesting findings based on exploratory data analysis, statistical tests, and predictive modelling that help to understand the main predictors and predictive variables related to PCOS.

Feature Influence and Correlation Patterns

Correlation analysis and feature importance rankings of Random Forest, XGBoost and CatBoost indicate that there are strong associations between PCOS prevalence and the features of follicle counts (left and right), cycle abnormalities, hair growth, and AMH levels. It is significant to note that follicle counts, and irregular menstrual cycles emerged as the most significant predictors, which is consistent with clinical evidence regarding increased antral follicle counts and ovulatory dysfunction in PCOS patients. This is consistent with clinical data on the enhancement of antral follicle counts and the failure to ovulate in PCOS patients. The FSH/LH ratio, which is usually less than 1 in PCOS cases, is also an important biomarker which indicates the problems in the hypothalamic-pituitary-ovarian axis. The clinical relevance of these features is further shown by SHAP analysis, which gives interpretable information on their effect on the model predictions.

Socio-Lifestyle Indicators

Binary lifestyle features such as hair growth, skin darkening, fast food consumption, and irregular exercise exhibited mild to moderate associations with PCOS. These features, though not dominant when in isolation, contribute considerably when combined with hormonal and physical indicators, creating risk profiles. This implies that lifestyle has a multidimensional effect on the development of PCOS.

Model Behaviour and Class Separation

The collection of models like Logistic Regression, Random Forest, XGBoost, CatBoost, and a Voting Classifier shows a good level of performance. The most accurate of them is CatBoost (90.18%, AUC-ROC: 0.942), followed by Random Forest (88.34%, AUC-ROC: 0.938). State-of-the-art methods of oversampling (SMOTE, ADASYN, Borderline-SMOTE) also showed increased sensitivity, which was beneficial to detect the minority PCOS-positive class (e.g., CatBoost-recall: 0.83). Confusion matrices of all models demonstrate high true-negative rates with CatBoost and Voting Classifier optimising the true-positive detection, which is critical in reducing underdiagnosis risks in the clinical environment.

Dimensional Reduction and Visualisation

PCA visualisation illustrates distinct clustering between PCOS and non-PCOS people, which is attributed to hormonal (e.g., LH, AMH) and physical (e.g., BMI, Waist-to-Hip Ratio) characteristics. This proves the fact that there are latent patterns distinguished between affected people. Further, SHAP summary and waterfall plots give a granular level understanding of the feature values, better interpretation of the model, and consistency between predictions and clinical reasoning.

Clinical Relevance and Applications

These lessons are quite practical for healthcare. The models make it possible to screen for PCOS at a very early stage, before the appearance of evident symptoms, using the already known clinical tests (e.g., AMH, follicle counts, BMI) and questionnaires (e.g., hair growth, weight gain) that are easier to complete. The value of feature engineering in medical data science is highlighted in the derived features like FSH/LSH ratio and BMI x Cycle length. The ensemble strategy ensures resource-limited prediction, and the SHAP analysis helps to build more clinical confidence in the effect of the features. The proposed framework enables cost-efficient and scalable diagnostic tools and the shift from reactive to proactive responses to symptom management and to the individualised approach to care.

CONCLUSION

The current research can be viewed as an effective data-driven model of predicting and analysing Polycystic Ovary Syndrome (PCOS) with the aid of advanced machine learning and statistical methods. The study used a real-life clinical dataset of 541 records and 44 attributes to investigate the important physiological, hormonal and lifestyle factors that influence the diagnosis of PCOS. The methodology involved careful preprocessing, feature engineering such as FSH/LH ratio, BMI x Cycle Length, model development to train a collection of models that include Logistic Regression, Random Forest, XGBoost, CatBoost and an ensemble Voting Classifier with high performance. It is important to note that CatBoost obtained the best accuracy of 90.18 and an AUC-ROC score of 0.942, while the ensemble model obtained an AUC-ROC score of 0.944. The most significant predictors, such as follicle counts, cycle abnormalities, and AMH levels, were consistent with the clinical literature, thereby strengthening their diagnostic value. Advanced oversampling methods (SMOTE, ADASYN, Borderline-SMOTE) were effective in handling the imbalance in classes and improved the sensitivity (e.g., CatBoost recall: 0.83) of PCOS-positive cases. The SHAP analysis was integrated to provide feature contribution and to predict in line with clinical rationale. The present research emphasises the synergy between domain knowledge and advanced data science with explainable AI for delivering valuable healthcare insights. This designed pipeline can be upscaled, adjusted and implemented in clinical decision support systems to enable the detection of PCOS at an early stage and accurately.

Future Works

The results obtained are promising, but there are opportunities for further studies:

- **Integration of Temporal and Longitudinal Data:** Future models could take advantage of temporal data (e.g., menstrual cycle records for a month or year) to get the temporal aspects of PCOS more effectively.
- **Explainability and Interpretability:** Despite the application of SHAP analysis, there is scope for considering more advanced explainable AI systems, such as LIME or attention-based mechanisms, for improved model transparency.
- **Cross-Population Generalisation:** The generalizability or applicability and utility of the model could be enhanced by testing the model using geographic and ethnic diversity in the sample of the population.
- **Real-Time Clinical Integration:** Incorporation of these models into electronic health records or index systems would facilitate better clinical workflow and decision-making in real-time systems
- **Multi-Omics Data Integration:** To identify the field biomarkers, the systems can include genetics, proteomic or metabolomic data. This will improve the personalised and dependable PCOS diagnoses.

The current study enables affordable and personalised access to reproductive health diagnosis for women. The models will become shorter, collated continuously according to the data and permit the machine learning model data to proceed to augment it as a tool committed to the emphasis of the focus of the population health as a pathway that has proven to be a significant issue for the people.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support provided by the STME SVKM's NMIMS Indore, Department of Computing, Coimbatore Institute of Technology and Symbiosis Institute of Computer Studies and Research, Pune. No external funding was received for conducting this research.

REFERENCES

Ahmad, R., Maghrabi, L. A., Khaja, I. A., Maghrabi, L. A., & Ahmad, M. (2024). SMOTE-based automated PCOS prediction using lightweight deep learning models. *Diagnostics, 14*(19), Article 2225. https://doi.org/10.3390/diagnostics14192225

Alagumariappan, P., Sathyamoorthy, M., Dhanaraj, R. K., Kannan, K., Kuppusamy, R., Dhanapal, R., Pamucar, D., & Shankar, N. (2025). Optimised hybrid machine learning framework for early diabetes prediction using electrogastrograms. *Scientific Reports, 15*, Article 8875. https://doi.org/10.1038/s41598-025-93495-3

Azziz, R., Carmina, E., Chen, Z., Dunaif, A., Laven, J. S. E., Legro, R. S., Lizneva, D., Natterson-Horowitz, B., Teede, H. J., & Yildiz, B. O. (2016). Polycystic ovary syndrome. *Nature Reviews Disease Primers, 2*, Article 16057. https://doi.org/10.1038/nrdp.2016.57

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Dhanaraj, R. K., Paramasivam, A., Vijayalakshmi, S., Emmanuel, C., Pallavi, P., Metkewar, P. S., & Ashwin, M. (2023). Design and implementation of an IoT-based smart digestive health monitoring device for the identification of digestive conditions. *Engineering Proceedings, 58*(1), Article 33. https://doi.org/10.3390/ecsa-10-16253

Hdaib, D., Almajali, N., Alquran, H., Mustafa, W. A., Al-Azzawi, W., & Alkhayyat, A. (2022). Detection of polycystic ovary syndrome (PCOS) using machine learning algorithms. In *2022 5th International Conference on Engineering Technology and Its Applications (IICETA)* (pp. 532-536). IEEE. https://doi.org/10.1109/IICETA53384.2022.9888677

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2008)* (pp. 1322-1328). IEEE. https://doi.org/10.1109/IJCNN.2008.4633969

Kailasanathan, N., Ezhilarasan, G., Selvarajan, S., Dhanaraj, R. K., Pamucar, D., & Shankar, N. (2025). Heart disease prediction with a feature-sensitised interpretable framework for the Internet of Medical Things sensors. *Frontiers in Digital Health, 7*, Article 1612915. https://doi.org/10.3389/fdgth.2025.1612915

Karthik, Y., Sruthi, R., & Sujithra, M. (2024). Polycystic ovary syndrome prediction through CNN-based image analysis: A deep learning-based approach. In *Proceedings of the 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)* (pp. 1547-1553). IEEE. https://doi.org/10.1109/I-SMAC63188.2024.10714821

Khanna, V. V., Chadaga, K., Sampathila, N., Prabhu, S., Bhandage, V., & Hegde, G. K. (2023). A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. *Applied System Innovation, 6*(2), Article 32. https://doi.org/10.3390/asi6020032

Krishna, S. (2023). *Identify PCOD for women in India*. Gynecology.Hyd Help India. https://doi.org/10.21428/e339930a.4e664f9f

Kumar, P., Chauhan, S., & Awasthi, L. (2023). Artificial intelligence in healthcare: Review, ethics, trust challenges and future research directions. *Engineering Applications of Artificial Intelligence, 120*, Article 105894. https://doi.org/10.1016/j.engappai.2023.105894

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science, 17*(3), Article e70056. https://doi.org/10.1111/cts.70056

Prabu, M. M., Dharshini, S., Dhivya, M., Dinesh Kumaran, J., & Emaya, M. (2025). Revolutionising PCOD detection with early machine learning methods. *International Journal of Scientific Research in Science, Engineering and Technology*. https://doi.org/10.32628/ijsrset251234

Raut, K., Katkar, C., & Itkar, S. A. (2022). Predictive modeling for PCOS using machine learning techniques. *International Research Journal of Engineering and Technology, 9*(1), 3710-3715.

Sidey-Gibbons, J., & Sidey-Gibbons, C. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology, 19*, Article 64. https://doi.org/10.1186/s12874-019-0681-4

Suha, S. A., & Islam, M. N. (2022). An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. *Scientific Reports, 12*, Article 17452. https://doi.org/10.1038/s41598-022-21724-0

Sun, H., Li, J., & Zhu, X. (2025). A novel expandable borderline SMOTE over-sampling method for the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering, 37*, Article 218399. https://doi.org/10.1109/TKDE.2025.3544284

Teede, H. J., Misso, M. L., Costello, M. F., Dokras, A., Laven, J., Moran, L., Piltonen, T., & Norman, R. J. (2018). Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Fertility and Sterility, 110*(3), 364-379. https://doi.org/10.1016/j.fertnstert.2018.05.004

Van der Ham, K., Moolhuijsen, L. M. E., Brewer, K., Sisk, R., Dunaif, A., Laven, J. S. E., Louwers, Y. V., & Visser, J. A. (2024). Clustering identifies subtypes with different phenotypic characteristics in women with polycystic ovary syndrome. *The Journal of Clinical Endocrinology & Metabolism, 109*(12), 3096-3107. https://doi.org/10.1210/clinem/dgae298

Vedpathak, S. (n.d.). *PCOS dataset* [Data set]. Kaggle. Retrieved August 2, 2026, from https://www.kaggle.com/datasets/shreyasvedpathak/pcos-dataset

Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Frontiers in Digital Health, 6*, Article 1430245. https://doi.org/10.3389/fdgth.2024.1430245

Zad, S., Chen, X., Nasser, R., Albarqouni, L., & Alaa, A. M. (2024). Machine learning algorithms for predicting polycystic ovary syndrome using electronic health records. *Frontiers in Endocrinology, 15*, Article 1298628. https://doi.org/10.3389/fendo.2024.1298628

Zhao, B., Wen, L., Huang, Y., Fu, Y., Zhou, S., Liu, J., Liu, M., & Li, Y. (2025). A deep learning-based automatic recognition model for polycystic ovary ultrasound images. *Balkan Medical Journal, 42*(5), 419-428. https://doi.org/10.4274/balkanmedj.galenos.2025.2025-2-14