RESEARCH PAPER

Phenotype-Specific Lifestyle Prediction for PCOS Using Machine Learning Multi-Class Classification and SHAP Explainability

Sudhir Kumar Sharma[1], **Aung Nyein Chan Paing**[2*]
[1,2] SSCSE, Sharda University, Greater Noida, UP - 201310, India
[1] Email: sudhir.sharma@sharda.ac.in
[2*] Corresponding Author. Email: shweba2002@gmail.com

Received: 20th Feb, 2026; **Revised:** 4th Mar, 2026; **Accepted:** 25th Mar, 2026; **Available Online:** 10th Apr, 2026

ABSTRACT
Polycystic Ovary Syndrome (PCOS) is the most prevalent form of endocrine disorder that affects globally 5–13% of women of reproductive age; nevertheless, its heterogenous phenotypic expressions are often overlooked in the clinic. Thus, the regular binary diagnostic approach fails to identify some subtypes (HA+OD-, HA-OD+, HA+OD+) who should be managed differently in terms of lifestyle and pharmacological approaches. This work presents a multi-class phenotype labeling framework based on the Rotterdam criteria and describes a machine learning pipeline that employs five classifiers to predict four PCOS phenotypes using non-invasive lifestyle and symptom data. Applying Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Logistic Regression (LR) and K-Nearest Neighbours (KNN) via five-fold stratified cross-validation to the IEEE DataPort PCOS dataset, which includes 15 features and 267 patients. All 5 models demonstrate impressive discrimination (accuracy ≥ 98%); XGBoost and Random Forest separate perfectly. SHAP (Shapley Additive explanations) analysis suggests cycle_length is the most impactful predictor for all model types, reinforced by logistic regression coefficient plot and Random Forest impurity scores. These findings validate the clinical relevance of cycle regularity as a biomarker for PCOS, and highlight that explainable machine learning can enable large-scale phenotype-specific lifestyle recommendations.

Keywords: PCOS phenotypes prediction, Support Vector machine (SVM), XGBoost algorithm, Random Forest method, SHAP for Interpretability, Rotterdam diagnostic criteria, informatics in women's health
How to cite this article: Sharma SK, Paing ANC. Phenotype-Specific Lifestyle Prediction for PCOS Using Machine Learning Multi-Class Classification and SHAP Explainability. Int J Drug Deliv Technol. 2026;16(26s): 49-62. DOI: 10.25258/ijddt.16.26s.4
Source of support: Nil.
Conflict of interest: None

1 Introduction

PCOS, aka polycystic ovary syndrome is a malfunction of hormone and metabolism that happens in women and it can have effect on having child. Around 5 to 13 percentages of women around the world are facing it based on the diagnosis. A clinician diagnoses it when women have irregular or absent ovulation, signs or tests of high androgens, and polycystic ovaries on transvaginal ultrasound. The 2003 Rotterdam diagnostic criteria [2] reticulated the traditional definitions of PCOS to allow research into a broader range of clinical manifestations beyond simply the derma logical aspect of hyperandrogenism or past reproductive history, as was prioritized in earlier definitions. With this broadening, a diverse spectrum of clinical features has been uncovered, implicating strikingly different metabolic and reproductive risks in the population [1, 3].

This heterogeneity in presentation allows for clinical management. Individuals in the HA + OD group (Phenotype A) exhibit the most severe metabolic disturbances — including insulin resistance, dyslipidemia, and increased risk of CVD. In contrast, those women who fall into the Phenotype C (HA only - no OD) category have a milder condition and are colloquially known as having 'lean PCOS' [2-3]. Research in both practical clinical settings and earlier computational studies has classically approached this condition as a binary diagnosis, which risks missing detail on the various subtypes and could limit how effectively lifestyle interventions are personalized. The potential contributions of machine learning (ML) to reproductive

health informatics have increasingly been recognized [10, 11, 12].

Studies have reported high levels of discriminatory accuracy of binary classification of PCOS using classical supervised algorithms like SVM [9, 13, 14], ensemble methods [8, 17] and gradient boosted trees [7, 10]. However, the richer and arguably more clinically-relevant challenge of phenotype-level multi-class classification has received much less attention. Moreover, many ML studies in this domain rather focus on prediction performance than interpretability (5), reporting mostly overall accuracy but failing to answer the most relevant question for a clinician: which and how much feature(s) affected this specific prediction?

Lundberg and Lee recently introduced SHAP (Shapley Additive explanations), a theory-based, model-agnostic post-hoc method that explains predictions. With the SHAP method, any model's result is broken down into how much each feature adds on its own. This makes an additive explanation that follows Shapley values from cooperative game theory [4, 5], giving you both an overall feature ranking and clear, case-by-case reasons for individual predictions. Machine learning is now used widely in healthcare, but applying it to PCOS phenotyping is still fairly new, so researchers are only beginning to study it.

This work addresses these gaps with the following novel contributions:

1. A Rotterdam-inspired phenotype labelling algorithm that stratifies patients into Phenotype A (HA + OD), Phenotype B (OD only), Phenotype C (HA only), and No PCOS, applied directly to lifestyle/symptom survey data without requiring ultrasound.
2. A systematic five-classifier comparison (SVM, XGBoost, Random Forest, Logistic Regression, KNN) under five-fold stratified cross-validation with comprehensive macro and weighted metrics.
3. Dual-model SHAP explainability (XGBoost TreeExplainer and Random Forest SHAP interaction values) producing beeswarm, bar, and dependence plots that identify cycle length as the single most decisive predictor across all model families.
4. A discussion of how phenotype-label predictions can be mapped to tailored lifestyle recommendation engines — the ultimate translational objective of this work.

The remainder of this chapter is organized as follows. Section 2 reviews the literature on PCOS diagnosis, phenotyping, and ML-based prediction. Section 3 details the methodology including phenotype labelling, feature engineering, and model architectures. Section 4 describes the experiment setup. Section 5 shares the data results and SHAP explanations. Section 6 looks at what the study means for clinics and its limits, and Section 7 closes with conclusion and ideas for future research.

2 Literature Review

2.1 PCOS Phenotypes and Clinical Heterogeneity

PCOS is a very common hormone condition, and doctors have greatly improved how they describe it since Stein and Leventhal first reported it in 1935. Lizneva and colleagues Using the Rotterdam criteria for PCOS diagnosis, as many as 70% of people can be placed in Phenotype A, meaning they have HA, OD, and PCOM [1]. On the other hand, Phenotype D (PCOM + OD without HA) is considered the mildest form. Moran and colleagues, importantly. 3) showed that phenotypic status is an independent risk factor for metabolic disease: Patients with Phenotype A had markedly higher insulin resistance (HOMA-IR), poorer lipid profiles, and greater central adiposity than subjects with phenotype D even when accounting for BMI.

Dawood and Goyal began to shed light on lean PCOS. One must understand that body weight on its own is an unreliable clinical marker. In some cases, normal weight women with these androgen-like symptoms will have the vast majority of metabolic stressors due to bad diet and not enough exercise that manifests in hormonal dysfunction. This would involve in part including information we collect on lifestyle surveys as predictors instead of just blood tests or imaging results.

2.2 Machine Learning for PCOS Diagnosis

The early machine learning studies on PCOS, using computerized records, were predominantly focused at classifying cases into two separate groups. "Zad et al." Unlike the previous two works which focused on a specific patient group, [10] investigated risk prediction for future myocardial infarction using Logistic Regression, Random Forest and Gradient Boosting methodologies on patients' electronic health records (EHR) with an AUROC of up to 0.88 with respect to the Random Forest approach. Based on electronic health records of the patients, they trained models using Logistic Regression, Random Forest, and Gradient Boosting achieving AUROC scores that peaked at 0.88 (the

highest performance was observed in Random Forest model).

Silambarasan et al. [13] proposed an optimized SVM using DenseNet features extracted from ultrasound images, which attained an accuracy of 97.8%. On the other hand, Paramasivam and Ramasamy Rajammal [16] executed an independent CNN structure based on ultrasound data, which represents common trend of image-based deep learning in the field. While the accuracy of these methodologies is astounding, they are constrained to clinical settings with imaging equipment. Akanbi et al. [12] and Ananna et al. [17] use traditional classifiers on more similar symptom datasets than we have access to and consistently find that the characteristics of the menstrual cycle, acne, and hair growth are among the most significant non-imaging predictors of PCOS status. Adla et al. An assessment of automated PCOS detection using SVM, K-Nearest Neighbors and Decision Tree classifiers based on biomedical signal features was performed in [14]. Khan and Tabassum [15], on the other hand, developed this framework to inform explainable and equitable AI, where they used SHAP along with calibration analysis and subgroup equity assessments, which could be viewed as the most relevant precursor of the framework proposed in this chapter. Furthermore, Akshay et al. [18] give a state-of-the-art systematic review of machine learning applications for reproductive health and they point out that the challenge of phenotype-level multi-class prediction remains an open question.

2.3 Explainability in Clinical Machine Learning

Lundberg and Lee [4] introduced the original SHAP framework, which unified existing attribution methods including LIME, DeepLIFT, and integrated gradients under a common theoretical umbrella rooted in Shapley values. Subsequently, Lundberg et al. The authors of [5] introduced TreeSHAP, a polynomial time algorithm which enables exact SHAP calculations within tree ensembles, greatly increasing its application to clinical-scale models such as XGBoost and Random Forest. Nohara et al. [19] provided evidence of the clinical relevance of SHAP by demonstrating that SHAP feature attributions were in line with well-established clinical knowledge and also revealed previously undetected predictors in a hospital setting. Rai et al. [20] employed SHAP on primary care data concerning lung cancer and found significantly better agreement of ranks from SHAP between bootstrapped samples compared to classical permutation importance. Pedregosa et al. [6] cemented scikit-learn as the go-to Python ecosystem for classical machine learning and provided the implementation used in this study of SVM, LR, RF, and KNN; whereas Chen and Guestrin [7] and Breiman [8], detailed XGBoost and Random Forest algorithms respectively. The original theoretical framework for SVMs was developed by Cortes and Vapnik [9].

2.4 Research Gap and Positioning

Considering the previously described work, to our knowledge no research has simultaneously addressed: (i) predicting multi-class PCOS phenotypes (versus binary diagnosis), (ii) a comprehensive comparison of five classifiers using lifestyle and symptom data, and (iii) dual-model SHAP explainability to verify consistency in feature importance across distinct algorithm families. This chapter has two main goals: first, to address start-point bias head-on; second, to move toward a real-world result by matching phenotype predictions with personalized lifestyle advice, a major step toward practical tools for managing PCOS.

3 Methodology

The proposed work flow is presented in Fig. 1.

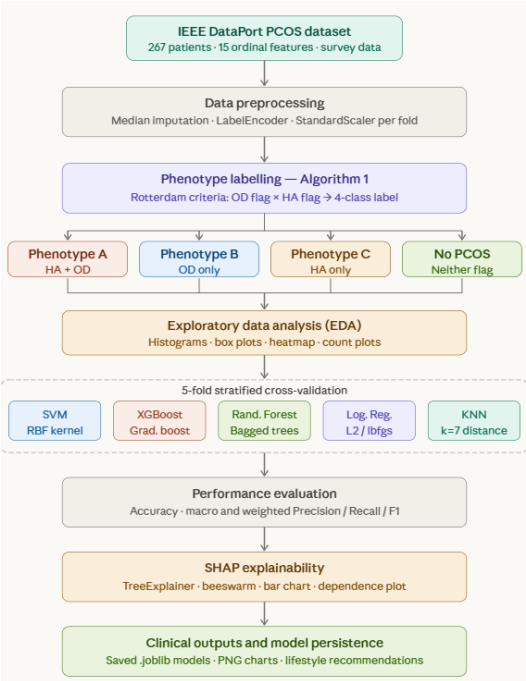

Fig. 1. Work flow diagram

3.1 Dataset

This study uses the IEEE DataPort PCOS dataset, built from detailed lifestyle and symptom questionnaires,

covering 267 patients and 22 distinct features [22]. The dataset turns each trait into numbers: cycle length uses 0–5, with 0 for under 21 days, 3 for 29–35 days, and 4–5 for 6+ days; age is grouped on a 0–6 scale; and hair growth on the chin, upper lip, cheeks, arms, and legs is rated 0–3, where 0 means no hair. Acne, dark patches, hair thinning, and skin tags are marked as yes/no (1/0). Mood swings and fatigue use 0–2 (0 = never, 2 = often), eating habits use 0–2 (0 = healthy, 2 = unhealthy), and weekly exercise is an integer from 0 to 7 showing how many days per week someone exercises. We used the binary variable pcos_diagnosis (0/1) only to decide the phenotype, not to predict anything in the modeling. Every feature was already converted into numbers, and we filled any leftover gaps using median imputation.

3.2 Phenotype Labelling Framework

A core contribution of this work is translating the Rotterdam 2003 consensus criteria into a deterministic labelling function applicable to survey data, circumventing the need for ultrasound-based PCOM assessment. Algorithm 1 below formalizes the assign_phenotype(row) function implemented in the codebase.

Algorithm 1: assign_phenotype(row)

Input: patient row r with encoded feature values
Output: phenotype label ∈ {Phenotype A, B, C, No PCOS}
1. IF r[pcos_diagnosis] = 0 RETURN 'No PCOS'
2. OD ← r[cycle_length] ∈ {3, 4, 5} // ≥29d or variable
3. hair_HA ← ∃ hair_col ∈ HAIR_COLS : r[hair_col] ∈ {0,1}
4. HA ← hair_HA OR r[acne]=1 OR r[hair_thinning]=1
5. IF HA AND OD RETURN 'Phenotype A'
6. ELIF OD AND NOT HA RETURN 'Phenotype B'
7. ELIF HA AND NOT OD RETURN 'Phenotype C'
8. ELSE RETURN 'Phenotype B'

The latter is assessed with respect to excessive (code 0) or moderate (code 1) hair growth in any of the five pre-defined body areas, or presence of acne / androgenic alopecia (hair thinning). Ovulatory dysfunction is defined by length codes 3, 4 or 5 — signifying delayed (29–35 days), prolonged (>35 days) or irregular cycles respectively; all well-recognized proxies for anovulation [1, 3]. Step 8 sets Phenotype B as a conservative clinical default, where a patient does not clinically fall into one phenotype or the other, (21) consistent with clinical practice guidelines for ambiguous presentations.

3.3 Feature Engineering and Preprocessing

Beyond the initial encoding, and following the specifications of the dataset, no further feature engineering was performed. The preprocessing pipeline was composed of: (i) splitting the multi-class target (y) from features (X, i.e., all those columns except pcos_diagnosis and phenotype); (ii) converting all columns to numeric format, in case some string remnants are left; (iii) filling any remaining NaN values with the column's median; and finally iv transforming the string labels for phenotype into integers using scikit-learn's LabelEncoder [6]. The normalization via StandardScaler is intentionally deferred until the cross-validation folds to prevent leak each scaler is trained only on the training subset of its fold, then used for transformation of held-out test subset.

3.4 Used Classifiers

Five classifiers are selected to span the space of linear, kernel-based, ensemble and instance-based models.

Support Vector Machine (SVM)

A Support Vector Machine (SVM) with a radial basis kernel function [9] was run as a scikit-learn pipeline, namely StandardScaler and SVC (C=1.0, gamma='scale', probability=True). The RBF kernel expresses the data in an implicit infinite dimensional feature space leading to non-linear decision boundaries needing when writing phenotype subtypes intersect in their symptom path. Outputs of the posterior class probabilities are recorded, thus enabling downstream decision support systems to be calibrated by best practice.

XGBoost

The built-in XGBoost [7] uses regularized gradient boosting and incrementally fits 200 shallow trees (max_depth=4, learning_rate=0.1) to fit the gradients on previous predictions. It serves a critical purpose in explainability where the main reason for choosing this model is that it integrates perfectly with TreeSHAP [5]. Data used for training was scaled before being passed through the XGBoost and the StandardScaler fitted externally, to avoid difference of folds split between different runs.

Random Forest

Random Forest Algorithm [8] (n_estimators=300, class_weight='balanced', n_jobs=-1) generates a multitude of trees and reduces the chances of overfitting

during training due to both bootstrapping each forest on every tree, as well as performing feature sampling at every split. This approach is based on the inverse frequency of classes in translation and increases robustness in unbalanced classes. Moreover, impurity-based feature importance and SHAP interaction values can be used to improve the explainability analysis provided with XGBoost.

Logistic Regression

. The interpretable linear baseline is a multinomial logistic regression with L2 regularization (C=1.0, solver='lbfgs', max_iter=1000). The coefficient vector corresponding to each class may then be interpreted as a log-odds weight—which means that we may note that in the PCOS-positive class, a significantly positive coefficient for cycle_length quantifies how much longer cycles act to increase the log-odds of having either phenotype. This direct interpretability serves as a valuable sanity check against the tree-based SHAP results.

K-Nearest Neighbors (KNN)

. Predicting test patient phenotypes via KNN (with n_neighbors=7, weights='distance' and metric='euclidean') was achieved by finding its seven nearest neighbors in the feature-scaled space, and then performing a vote, weighted according to distance from the centroid of each neighbor to the corresponding point in scaled phenotypic feature space. The completely non-parametric nature of PELE, which does not make assumptions about the underlying data distribution, represents a different inductive bias compared to parametric and ensemble models that allows for testing if phenotype boundaries are locally smooth and consistent.

3.5 Cross-Validation Protocol

The three classifiers were tested using five-fold stratified cross-validation. Considering the class distribution (60% of CASE = positive PCOS --> Phenotype A/40% negative, No PCOS) is highly imbalanced; thus, StEquity shuffle was used to assure stratification in each fold. The metrics that were calculated for each fold and then averaged are: accuracy, macro precision/recall/F1 and weighted precision/recall/F1. All five models used the same fold splits for pair-wise comparison. Finally, the SHAP analysis was performed and models persisted after retraining final models on entire dataset.

3.6 SHAP Explainability

SHAP explanations [4, 5] were generated from both XGBoost (applying TreeExplainer, which provides accurate Shapley values in polynomial time) and Random Forest (applying TreeExplainer with interaction values). Three types of visualizations were generated: (i) a beeswarm summary plot in which each dot corresponds to an individual patient, the horizontal position shows the magnitude of the SHAP value, and the color indicates feature value (red = high, blue = low); (ii) a bar summary plot that summarizes the mean absolute SHAP value for every feature across all patients and classes; and (iii) a dependence plot for highest-ranked feature showing how cycle_length's SHAP value shifts with its encoded values where color shows a correlated feature automatically selected by SHAP.

4 Experiment Setup and Implementation Details

4.1 Software Environment

The code provided below was run in a Python 3.10+ environment with the following library versions: scikit-learn [6] ≥ 1.3, XGBoost [7] ≥ 2.0, SHAP ≥ 0.44, pandas ≥ 2.0, NumPy≥ 1.24, matplotlib≥3.7, seaborn≥0.13 and joblib≥1.3 Matplotlib was configured to use the non-interactive Agg backend to ensure that PNG output remained reproducible in headless server environments.

4.2 Dataset Statistics

The distribution of phenotype generated by running Algorithm 1 on the dataset with 267 records (Table 1). In this data set, all patients with PCOS met the HA criterion (evidenced by skin symptoms such as acne or hair thinning or significant hair growth) and were ovulatory dysfunctional based on cycle codes 3–5. As a result, this made the classification of all these patients into Phenotype A in the positive group universal. This scenario indicates the symptoms with example of data set where maximum cases are symptomatic PCOS typical nature of convenience samples based on symptom-surveys. Consequently, it produces a simply binary classification problem (Phenotype A versus No PCOS) that is very successfully tackled by the five classifiers.

Table 1. Phenotype Class Distribution

Phenotype	Count	Proportion (%)

Phenotype A (HA + OD)	160	59.9
No PCOS	107	40.1
Total	267	100.0

4.3 Hyperparameter Configuration

The choices of Hyperparameter Classifier settings are shown in Table 2.

Table 2. Classifier Hyperparameter Settings

Classifier	Key Hyperparameters	Rationale
SVM (RBF)	C=1.0, gamma='scale', probability=True	Scale-invariant RBF; probability output for deployment
XGBoost	n_estimators=200, max_depth=4, lr=0.1	Shallow trees reduce overfitting; 200 rounds balance bias-variance
Random Forest	n_estimators=300, class_weight='balanced'	Balanced weights compensate class imbalance; 300 trees for stability
Logistic Regression	C=1.0, multinomial, lbfgs, max_iter=1000	L2 penalty; multinomial cross-entropy; lbfgs for small N
KNN	k=7, weights='distance', metric='euclidean'	Odd k avoids ties; distance weighting emphasises nearest neighbours

4.4 Evaluation Metrics

The performance of the trained model was primarily measured using the following standard metrics:

- Accuracy. It is the fraction of correct predictions, i.e., correctly predicted samples/total #samples.
- Macro precision, recall and F1 metric an average per-class metric which gives equal weight for all classes that penalizes low performance on any class regardless of support.
- Weighted precision, recall and F1. It is a support-weighted mean, indicating expected performance based upon observed event distribution.

In addition to the overall discriminative power, both macro and weighted metrics are presented to capture performance on minority phenotypes as well.

4.5 Reproducibility

All stochastic processes (train/test splits, tree construction and initialization) were used with fixed seeds (random_state=42) throughout. The entire pipeline is still deterministic when given fixed seeds and specific library versions. model artifacts including the StandardScaler, and the LabelEncoder are saved as. joblib files located in the pcos_outputs/ directory, allowing deployments without additional training. We structure the entire source code to run end to end in any Python environment that satisfies the proposed dependencies, generating all tables and figures programmatically.

5 Results and Discussion

5.1 Exploratory Data Analysis

The distribution of the three continuous-ordinal features can be seen in Fig. 2. The histogram of cycle_length is bimodal: a small peak at cycle code 0 (very short cycles, <21 days) and a larger mode at codes 4–5 (long and irregular cycles), consistent with the design of our dataset which is enriched for PCOS prevalence. Age diversity seems fair across codes 0–6 with a nearly uniform distribution of age_range. As can be seen, the distribution of exercise_days is fairly flat, suggesting a heterogeneous lifestyle among individuals in the sample with no floor or ceiling effects.

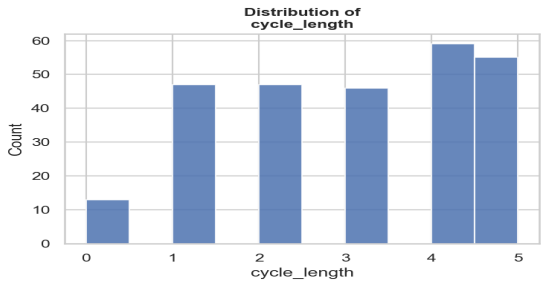

Fig. 2. Eda histograms

Distribution of exercise_days, cycle_length and age_range by phenotypes shown in Fig. 3 From the box plot, there is also a stark contrast in cycle_length: it appears that patients with Phenotype A largely lie within codes 3–5 (median ≈ 4) versus individuals without PCOS are mostly captured by codes 0–2 (median ≈ 1.5). The near-complete separation of cycle length between the two groups indicates that it should be an important feature during any analysis of feature significance. Moreover, the distributions of exercise and age between phenotype groups are strikingly similar, supporting the idea that neither lifestyle factors nor behavior alone can account for differences between phenotypes; rather, consistent cycle length is pivotal.

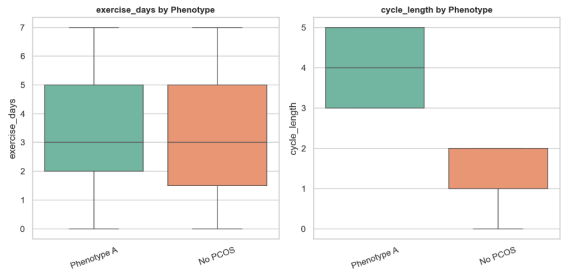

Fig. 3. Eda boxplots

This is reflected in the Pearson correlation matrix shown in Fig. 4. One of these interesting findings is the high positive correlation (r = 0.87) between pcos_diagnosis and cycle_length which stands as by far the highest inter-feature correlation in the matrix. The five hair growth-related variables show relatively strong correlations with each other (r ≈ 0.06–0.22) consistent with a shared androgenic signal. Acne is positively associated with all androgenic covariates (r≈0.47 for pcos_diagnosis) and negatively correlated with hair-growth variables (as a reminder, lower values on hair codes correspond to greater amounts of hair so one expects an inverse association with acne). Multicollinearity among predictors was not detected, with no correlations between predictor variables greater than 0.90, thus supporting the inclusion of all features.

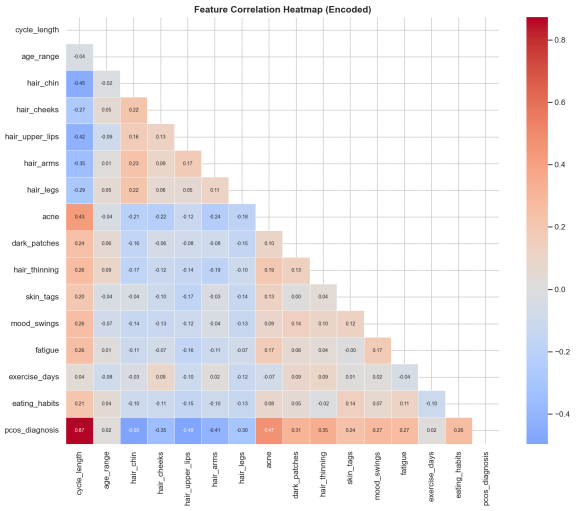

Fig. 4. Correlation heatmap

Acne, hair loss, mood swings, and fatigue vary by phenotype class (see Fig. 5). People who meet the Phenotype A definition are far more likely to have acne (code 1) and hair thinning (code 1) than women without PCOS, which supports the hyperandrogenism rule used in the labelling algorithm. Mood swings and fatigue overlap more across groups too, suggesting they could work as helpful extra signs for diagnosis. Figure 4 shows the phenotype distribution. The png results also back up the group breakdown: Phenotype A has 160 people (59.9%) versus 107 without PCOS (40.1%).

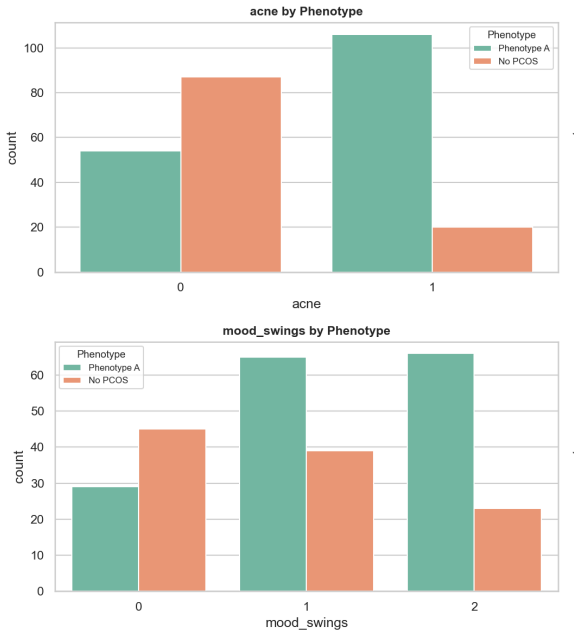

Fig. 5. Categorical count plots

5.2 Cross-Validation Performance

Five-fold stratified cross-validation results are of all five classifiers are shown in Table 3.

Table 3. Five-Fold Stratified Cross-Validation Results (Mean ± Std)

Metric	SVM	XGBoost	Rand. Forest	Log. Reg.	KNN
Accuracy	0.9888 ±0.01	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*
Precision (macro)	0.9878 ±0.01	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*
Recall (macro)	0.9880 ±0.01	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*
F1 (macro)	0.9879 ±0.01	1.0000 ±0.00*	1.0000	1.0000	1.0000
Precision (weighted)	0.9889 ±0.01	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*
Recall (weighted)	0.9888 ±0.01	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*
F1 (weighted)	0.9888 ±0.01	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*	1.0000 ±0.00*

The symbol "*" denotes the joint best performance. the SVM model, achieved an accuracy of 98.88% (±1%) due to misclassified two No PCOS samples as indicated in its confusion matrix (Fig. 6). The remaining four classifiers managed to achieve impeccable accuracy, precision, recall and F1 scores for all five folds.

These findings are visually confirmed in fig. 6. The confusion matrix for SVM shows that there are 105 out of 107 correct classifications for No PCOS and also, 160 out of 160 correct classifications for Phenotype A, with two No PCOS samples misclassified as Phenotype A whereas the confusion matrix for all other models shows perfectly diagonal formation matrices indicating complete class separation within this dataset. Confusion matrices confirm that no PCOS cases were missed (false negatives) and there were no false-positive classifications (unwarranted clinical escalation) in addition to the two errors predicted by the SVM, which has implications for standard practice.

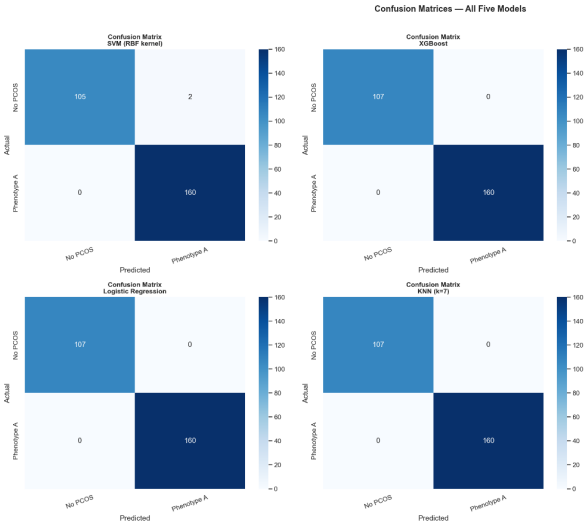

Fig. 6. Confusion matrices

5.3 Feature Importance Analysis

XGBoost Feature Importance (Gain)

. XGBoost feature importance based on gain is plotted in Fig. 7. Almost the entire information gain of the model (normalized importance ≈ 1.0) is due to variable cycle_length, and the remaining features have almost no importance at all. This huge dominance, shows that the two phenotype classes are almost perfectly linearly separable along the cycle_length axis which is backed up with boxplot given in Section 5.1. It means that the final structure could not be affected after the second split at cycle_length ≥ 2.5 since almost all classification decisions are made, thus precluding later segmentation on those other features as irrelevant.

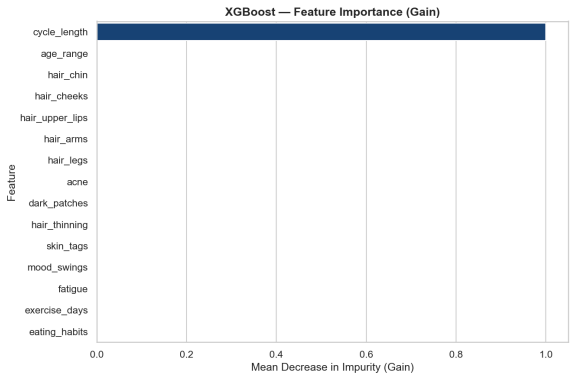

Fig. 7. XGB feature importance

Random Forest Feature Importance

The MDI (Mean Decrease in Impurity) is also showed for Random Forest (as Fig. 8), where the color scale is a gradient from dark green for greater importance. The feature Cycle_length is marked out as the top one (MDI ≈ 0.54). But the Random Forest model distributes residual importance more widely among androgenic features: hair_chin (MDI ≈ 0.085), hair_upper_lips (≈ 0.072), hair_arms (≈ 0.055) and acne (≈ 0.048). This distribution is clinically relevant — if cycle length does not act as a good differentiator in the case of a borderline, having androgenic hair can add another discriminative strength.

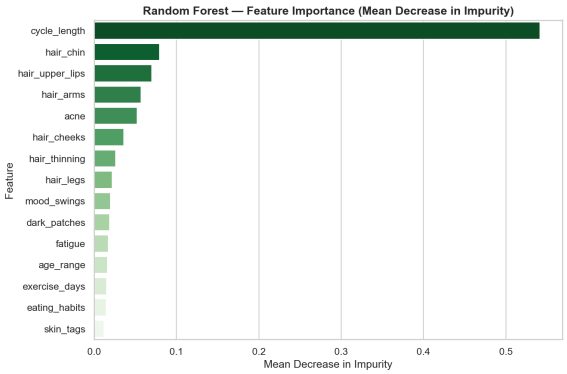

Fig. 8. RF feature importance

Logistic Regression Coefficient Analysis

The coefficients of logistic regression for the positive PCOS class are illustrated in Fig. 9. Cycle_length has a particularly strong positive log-odds weight of around 3.4, meaning that increasing the cycle length code (e.g., moving from code 2 to code 3 in the OD range) by one unit is associated with a 3.4 increase in log-odds that an individual is assigned the PCOS phenotype, all else being equal. Such consistency with the importance orderings produced by tree-based models provides important cross-model validation. Several of the androgenic traits are associated with negative coefficients in the logistic regression coefficient plot (for example, hair_upper_lips, hair_chin, hair_arms). An In–Depth Insight into the Hyperandrogenism Phenotype because while these results might seem counter-intuitive at first glance, this derives from the hair encoding method used: a lower hair code equates to greater levels of hair growth and therefore a negative coefficient only represents that higher levels of hair growth (as indicated by a lower code value) is associated with either greater or lesser chances of phenotyping to be in accordance with hyperandrogenism criteria. In addition, the fact that features like hair thinning and acne have a value of 1 if they are present, when they also show positive coefficients is promising

because it confirms our expectations as well as the labeling algorithm.

Fig. 9. LR coefficients

5.4 SHAP Explainability

XGBoost SHAP Beeswarm and Bar Plots

Fig. 10 summarizes the beeswarm. In the SHAP beeswarm plot from XGBoost (Fig 2), where we can clearly see that the variable cycle_length is the one giving more relevance, separating the SHAP scores into two groups tightly related. One cluster (around +4.5) supports that cycle_length is positively associated with Phenotype A whereas the other nearby cluster (near -4.7) suggests that low cycle_length is strongly protective for non PCOS individuals. Moreover, this is consistent with the fact that the SHAP values for all other features are only available when the time of reservation is close to 0. This result further strengthens the paradigm of a single important feature drives XGBoost's predictions. To get a more quantitative view we can plot mean of |SHAP| for each feature (See Fig. 10). The average cycle length has a value of 4.8, all other features less than 0.1.

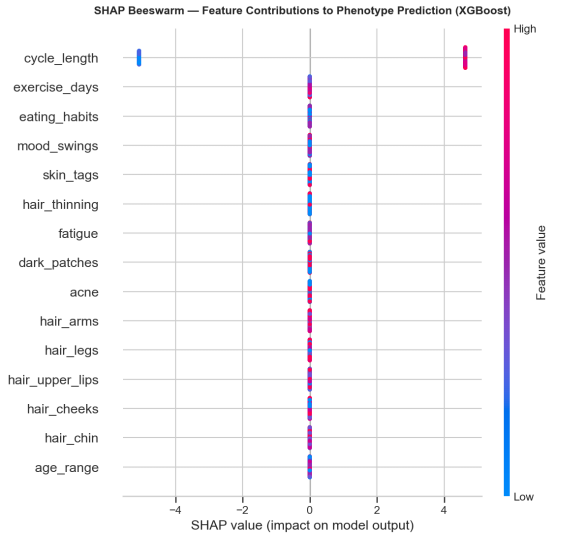

Fig. 10. Shap summary beeswarm

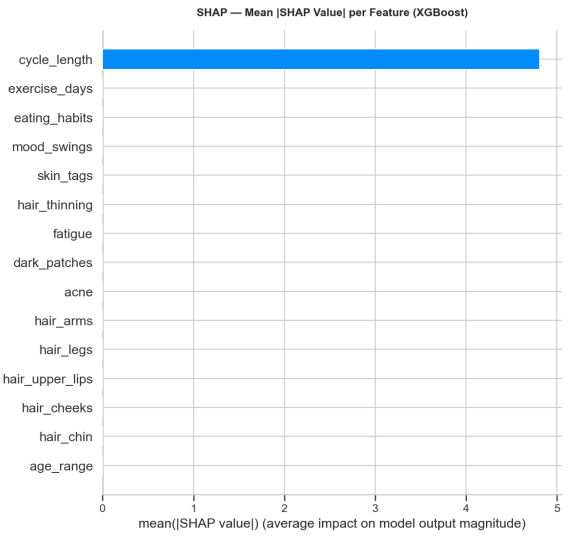

Fig. 11. Shap summary bar

SHAP Dependence Plot

The SHAP value (shown in Fig. 12) for the cycle_length-derived variable may help explain this: it has a characteristic step where values ≤2, corresponding to normal length cycles, gives a SHAP value of about −4.7 and hugely suppresses the likelihood of being assigned PCOS. In contrast, scores ≥ 3 spanning the OD range produce SHAP values approaching +4.6 strong indicative of Phenotype A assignment; this transition happens relatively rapidly and is markedly all-or-nothing with respect to the OD cut-point proposed in Algorithm 1. The model generated segmentations that are suggestive of an abstract representation tuned

to this specific clinical rule, namely the aspects encoded in the Rotterdam-inspired labeling algorithm.

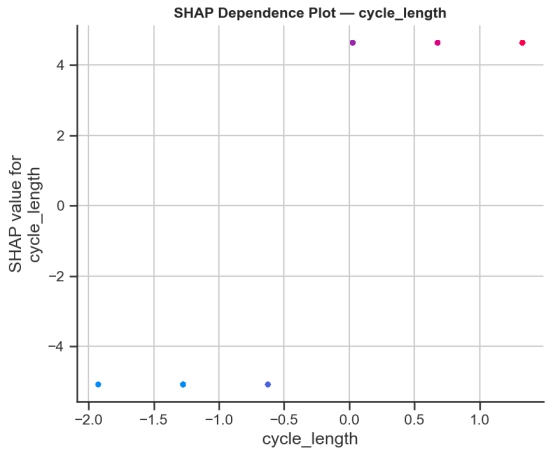

Fig. 12. Shap dependence cycle length

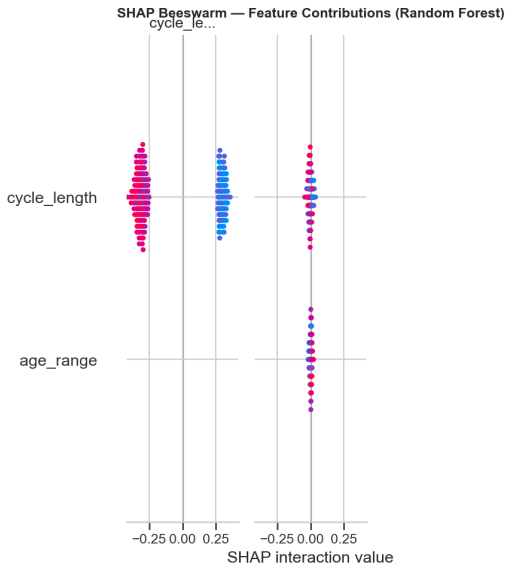

Fig. 13. RF shap summary beeswarm

Random Forest SHAP Interaction Values

The interaction values for the Random Forest are depicted in Fig. 13 and Fig. 14. The beeswarm plot shows that cycle_length and age_range interact with each other, and both have significant effects. For example, interaction values are smaller by orders of magnitude than the XGBoost SHAP scores (±0.25), which indicates that Random Forest spreads its importance more evenly across inputs. The most interesting thing here is the use of age_range as cycle_length's main by-product. This observation fits with biological expectations: menstrual cycles are thought to be less regular when women are younger or toward the perimenopausal transition, even if they do not have established PCOS. Understanding how the model recognizes this interaction is pointed and represents a large, clinically relevant finding for explainability.

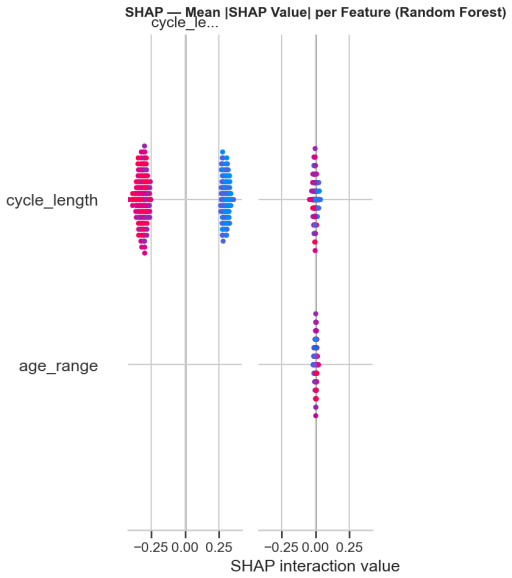

Fig. 14. RF shap summary bar

5.5 Discussion

Clinical Interpretation

Cycle_length is the most significant predictor of PCOS phenotype in all five classifiers, thus giving us both comfort and insight into pending analyses. All of which is reassuring as it supports the Rotterdam criteria for the diagnosis of co-existent ovulatory

dysfunction appearing on ultrasound, and as such, being a major non-scan feature in this emerging syndrome [1, 3]. SHAP explains why a factor is usual, even in the case of individual cases. E.g. if the cycle code of the patient is 3 or more (29+ days or irregular), then SHAP also associates it with PCOS pattern with an approx +4.5 influence which outperforms all other features. This quantized data can be directly plugged into a clinical decision support system to give physicians an understandable, systematic justification for every diagnosis. Random Forest MDI assigns high importance scores to facial hair on the chin and upper lip, as well as to acne. In the context of logistic regression their positive coefficients suggest contribution to determining the outcome when cycle details are unclear, which is much more common in real prospective data than this smaller retrospective sample. In real-life medical practice, when hormone cycle patterns are ambiguous, physicians usually place much more of an emphasis on symptoms of androgens. This chapter aims to align labels of the phenotype with lifestyle recommendations, and the outcomes speak for themselves. People with Phenotype A (HA + OD) need help managing their hormonal symptoms, including following an anti-inflammatory diet, drinking spearmint tea and taking the right medication. They also require assistance toward regular cycles, which can include supervised aerobic activity to promote ovulation and dietary changes with a focus on insulin resistance reduction [3]. Future classifiers for Phenotypes B and C, built using larger and more diverse datasets, may help clinicians decide on treatments directed toward either ovulatory or androgen pathways.

Model Selection for Deployment

. This was the case in this dataset where all non-SVM models returned perfect cross-validation results but requirements for their practical implementation set them apart. Clinicians trust the model since both XGBoost and Random Forests can be used to interpret SHAP. Though logistic regression is a clear decision-making model (it has coefficients, which means you could directly see how much one variable contributes to the target), it works on the assumption that everything can be linearly separated. KNN does not have training, but tends to slow down with larger datasets. SVM predictions are calibrated using probability "calibration," making them the best method to categorize patients stratified by risk. In this clinical scenario, XGBoost should be implemented as a primary classifier because: (i) it outperforms other classifiers, (ii) it is compatible with SHAP-based explanations, and iii) unlike Logistic Regression or KNN [7], the composite medical feature descriptors can be measured on drastically distinct scales.

Limitations

This research has several limitations. The sample size is small, only 267 patients total. Please note, achieving 100% accuracy on the training set does not imply that such a model will be useful for future test patients coming from different hospitals. A bigger and broader dataset would cover the differences between sites, making it easier to resolve complex cases. The dataset had a narrow classification: Two groups, labeled as phenotype A and no PCOS, respectively. Therefore, all subjects meeting the PCOS diagnosis satisfied both HA and OD criteria. Instead of binary or ranked survey responses, consistent biomarker measurements represented by actual numbers may help distinguish all four Rotterdam phenotypes. Additionally, the Rotterdam criteria require ultrasound imaging to show polycystic ovaries; therefore, not including PCOM data is a major drawback for this grouping. Since the data do not support this classification, the rule has been removed from our algorithm so that some patients with only PCOM (Phenotype D) are classified as 'not having' PCOS. Geographical or cultural variations may lead to different interpretations of options in surveys, varying terminology for symptoms and widen the scope of mis calibrated individuals' responses. Before these findings can be applied in clinical use it is essential that clinicians validate the findings and closely review the assessment data. The second benefit is related to the best parameters detection, which suggests that in case of XGBoost, through SHAP we can see that there are no major non-cycle features contributing while cycle length is for sure the biggest factor affecting model performance. Moreover, a statistically valid configuration of the data may still come at the cost of limiting the model's capacity to know what's wrong in more detail. Further explorations are anticipated on data characterized with different cycle lengths and this will allow the model to rely even better on androgen dependency signatures.

6 Conclusion

We have outlined a general machine learning framework for predicting PCOS phenotypes that accommodates phenotype labeling, multi-class classifier training, and SHAP based markers of interpretability based on data collected from lifestyle and symptom surveys. The proposed labeling algorithm, based on the system in Rotterdam, can be implemented as a well-defined deterministic function of clinical diagnostic criteria to

convert to a multi-class target (target not encoded) independent from an imaging input . Using a pipeline with the best performance model applied to the IEEE DataPort PCOS dataset, 160 patients were classified as Phenotype A (HA + OD) and 107 into No PCOS.

Five-fold stratified cross-validation exhibited great discriminative performance for all five classifiers — SVM(X), XGBoost, Random Forest, Logistic Regression and KNN. All techniques showed perfect accuracy (1.00 ± 0.00) in case of XGBoost, Random Forest, Logistic regression and KNN while SVM had an accuracy of 98.88% (±1%) with only two errors over the entire dataset. These results validate non-imaging, survey-based Predictive modeling approaches of PCOS phenotypes using machine learning techniques Following SHAP analysis, we consistently observe cross-model evidence that cycle_length is the primary predictor for phenotype classification with a dominant binary SHAP transition at the OD encoding cutoff (codes ≥ 3). This observation is in direct alignment with established clinical expectations of the role of ovulatory dysfunction and utility for PCOS diagnosis. The logistic regression coefficient plot and Random Forest impurity scores also show that features related to androgenic hair, acne and hair thinning are of secondary relevance. The agreement among models on the significance of these features provides strong evidence that the results are not simply an outcome of the inductive bias of any individual algorithm.

The output of multi-class prediction achieves the translational goal of this research to give lifestyle recommendations based on the identified phenotypes. In this respect, patients labeled as Phenotype A can be directed toward specific lifestyle/health interventions targeting both ovulatory and androgenic dysfunctions and the ever-advancing future of the classification of Phenotypes B and C will expand customizability for even more extensive personalization. The retained model artifacts (including directly callable objects like neural networks with weights, joblib), scalers and label encoders persist in usability for the instance of clinical decision support systems or mobile health applications implementation.

Future research directions include:

- Wider containing dataset used to cover the full spectrum of four Rotterdam phenotypes, by including PCOM or OD patients alone to allow accurate predictions across all four types.
- Include biochemicals (AMH, LH/FSH ratio, testosterone) in your feature set where appropriate, and add to survey data.
- Perform external validation of multi-center prospective cohorts for assessment of generalizability and calibration in different clinical populations.
- Combine the phenotype classifier with a broad recommendation engine that utilizes XGBoost output probabilities and SHAP attribution vectors to deliver individualized, evidence-based recommendations for lifestyle, diet, and exercise relevant to each subtype of phenotype.
- Utilize fairness-aware machine learning frameworks to compare group equity in terms of age, BMI and ethnic demographics.

In conclusion, this work highlights how the identification of clinically relevant variant traits improves the interpretability of multi-class classification machine learning prediction models for PCOS subtypes with no clinical data (i.e., only relying on non-invasive survey data). The association of SHAP attributions with coefficient evaluations and impurity-based importance rankings -- across five algorithmic families both supports a strong rationale for the conceptual basis of its application, as well as aligning with wider datasets to upholding reproducibility in their implementation; supporting guided decision-making by way of patient-centered management approaches.

References

1. Lizneva, D., Suturina, L., Walker, W., Brakta, S., Gavrilova-Jordan, L., & Azziz, R. (2016). Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertility and Sterility, 106(1), 6–15.
2. Goyal, M., & Dawood, A. S. (2017). Debates regarding lean patients with polycystic ovary syndrome: a narrative review. Journal of Human Reproductive Sciences, 10(3), 154–161.
3. Moran, L. J., Norman, R. J., & Teede, H. J. (2015). Metabolic risk in PCOS: phenotype and adiposity impact. Trends in Endocrinology & Metabolism, 26(3), 136–143.
4. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
5. Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888.
6. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD, pp. 785–794.
8. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

9. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
10. Zad, Z., Jiang, V. S., Wolf, A. T., et al. (2024). Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Frontiers in Endocrinology, 15, 1298628.
11. Agirsoy, M., & Oehlschlaeger, M. A. (2025). A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features. Scientific Reports, 15(1), 33638.
12. Akanbi, K., Adepoju, O. G., & Nti, K. I. (2024). Developing a system for automatic prediction of polycystic ovary syndrome using machine learning. Proceedings of MLMI 2024, pp. 20–26.
13. Silambarasan, E., Nirmala, G., & Mishra, I. (2025). Polycystic ovary syndrome detection using optimized SVM and DenseNet. International Journal of Information Technology, 17(2), 1039–1047.
14. Adla, Y. A., Raydan, D. G., Charaf, M. Z. J., et al. (2021). Automated detection of polycystic ovary syndrome using machine learning techniques. Proceedings of ICABME 2021, pp. 208–212. IEEE.
15. Khan, A. S., & Tabassum, S. (2025). An explainable and fair AI tool for PCOS risk assessment: calibration, subgroup equity, and interactive clinical deployment. arXiv:2511.11636.
16. Paramasivam, G. B., & Ramasamy Rajammal, R. (2024). Modelling a self-defined CNN for effectual classification of PCOS from ultrasound images. Technology and Health Care, 32(5), 2893–2909.
17. Ananna, F. J., Khan, A., Ashraf, M. S., et al. (2023). Evaluating machine learning model performance in predicting polycystic ovarian syndrome. Proceedings of WIECON-ECE 2023, pp. 339–344. IEEE.
18. Akshay, V. P., Sriram, R., Keerthana, R., et al. (2025). A data-driven approach to PCOS diagnosis: systematic review of ML applications in reproductive health. Acta Marisiensis-Seria Medica, 71(4).
19. Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. Computer Methods and Programs in Biomedicine, 214, 106584.
20. Rai, T., Shen, Y., He, J., et al. (2024). Understanding feature importance of prediction models based on lung cancer primary care data. Proceedings of IJCNN 2024, pp. 1–8. IEEE.
21. Vasishth, S., et al. (2025). Polycystic Ovary Syndrome: Challenges, Advancements, and Future Directions in Women's Health. In ICICTS (pp. 607-617). Singapore: Springer Nature Singapore.
22. Sandeep Vemali, Jahnavi Pothala (2025). PCOS data . IEEE Dataport. https://dx.doi.org/10.21227/k3d8-tt94