Mental Health Correlates of PCOD in Indian Women: A Machine Learning-Based Predictive Approach

Ria Pyne[1], *Arundhati* Roy[1], *Ahona* Ghosh[2], *Kamolika Sarkar* Puja[1], *Biswajit* Chakraborty[2], and *Saikat* Das[1]

[1]Brainware University, Kolkata, India
[2]Department of Computer Science and Engineering (IoT, CS, BT),
University of Engineering and Management, Kolkata, India

Abstract. Polycystic Ovarian Disease (PCOD) is a multifactorial endocrine disorder that significantly affects reproductive-aged women, manifesting through a combination of physiological and psychological symptoms. In South India, cultural expectations, limited awareness in rural areas, and the stigma surrounding mental health often compound the emotional burden of the disorder. This chapter explores the intersection between PCOD and mental health, analyzing patterns of fatigue, mood swings, stress, cognitive challenges, and lifestyle changes in diagnosed women. Polycystic Ovarian Disease (PCOD) is a multifactorial endocrine disorder affecting reproductive-aged women, often leading to infertility, obesity, and psychological distress. The growing application of machine learning in healthcare provides promising avenues for accurate, early detection of PCOD. In this chapter, we investigate the efficacy of gradient boosting algorithms for PCOD prediction using a dataset containing physiological and psychological attributes. Employing train-test splits (80-20, 50-50, 66-34) and 10-fold cross-validation for robustness, our model achieves a commendable accuracy of **97.65%**, highlighting its potential utility for clinical decision support systems. By combining traditional survey-based analysis with advanced predictive models, the study emphasizes the critical link between PCOD symptoms and mental health outcomes. The findings support the integration of mental wellness indicators in diagnostic frameworks and advocate for culturally sensitive healthcare interventions across southern Indian populations.

1 Introduction

Polycystic Ovarian Disease (PCOD) or Polycystic Ovary Syndrome (PCOS) is a complex hormonal disorder that occurs in a significant percentage of women in their reproductive life. It is described with the imbalance of the normal female sex hormones and may cause various clinical manifestations in diverse degrees and combinations among individuals. The most frequent manifestations are irregular periods of menstruation, excessive growth of body hair,

persistent acne, sudden changes of body weight, and the appearance of numerous cysts in the ovaries. In addition to the reproductive implications, PCOD has a far-reaching systemic effect that spills out to the metabolic and psychological realms. It has often been linked with type 2 diabetes, insulin resistance, high blood pressure and increased levels of cholesterol, therefore increasing the chances of cardiovascular disease in the long-term. These physical symptoms are merely a tip of the iceberg since the disorder has equally a heavy psychological toll to the affected women, which is reflected in the form of stress, anxiety, depression, social withdrawal, and low quality of life.

The symptoms of PCOD have a very dynamic nature and are likely to be diverse, which makes the process of diagnosis of the disease quite complex and subtle. The diagnosis usually depends on a set of physical tests, ultrasound results, and hormonal tests, including the determination of luteinizing hormone (LH), follicle-stimulating hormone (FSH), and the levels of testosterone. Nonetheless, these tests cannot be conclusive and sometimes they can fail to capture mild and unusual manifestations of the disease. In other situations, women might be psychologically distressed before the emergence of classical cystic images on ultrasound, resulting in late or misdiagnoses. Therefore, swift, accurate and extensive diagnosis is necessary to prevent short and long-term complications so that medical control, dietary alteration and reshaping of lifestyle can be taken in time.

The past few years have seen a growing interest in computational and data-driven methods among researchers and clinicians in order to enhance the accuracy of diagnosis and the prediction of diseases. Machine learning algorithms can utilize large datasets with multidimensional features, can learn upon the past patient data, and can create predictive models that may help clinicians to make decisions. The Gradient Boosting Classifier among these algorithms has received a lot of attention because it is capable of refining weak predictive models repeatedly into a strong ensemble with high levels of accuracy and generalizability.

In this chapter, the authors examine how a Gradient Boosting Classifier could be used to predict PCOD using an integrated dataset that combined physiological and psychological features. Unlike conventional diagnostic research that uses mostly medical predictors, the self-reported psychological and lifestyle measures used here include fatigue, abdominal pain, sleep patterns, emotional distress, and fertility concerns. The addition of these non-clinical aspects to the model improves its predictive ability and makes it more realistic in terms of diagnostic processes in the real world, where patient experience and mental health can equally inform clinical judgments as laboratory findings.

Among the most important impetuses to conduct the study is the fact that mental health disruptions, including chronic stress, anxiety, and depression are linked to the symptoms of PCOD. The physiological explanation of this relationship is the existence of a hypothalamic-pituitary-adrenal (HPA) axis, a system that controls the reaction to stress and hormonal activity. Cortisol dysregulation is a result of prolonged psychological stress, which in its turn disrupts the levels of insulin and androgens, worsening the hormonal imbalance inherent in PCOD. With the help of several strategies, the proposed machine learning model was tested to guarantee robustness, reliability, and generalization. The data was split into multiple training and testing sets in 80-20, 66-34, and 50-50 ratios. The Gradient Boosting Classifier portrayed high rates of accuracy of over 97 percent, confirming that it could identify clinical and psychological predictors of PCOD with high precision.

The analysis on feature importance showed that stress score, emotional stability index, body mass index (BMI), and menstrual regularity were found to be some of the strongest predictors of PCOD. This observation supports the thesis that psychological health cannot be separated from physical health, especially when it comes to women's hormonal disorders. This study also introduces a significant sociocultural aspect because of the South Indian context. The urbanization process and dietary pattern changes, along with increased professional

requirements, have helped to raise the level of stress and sedentary lifestyle among women, particularly those between their 20s and 30s.

Finally, the results of this paper promote a shift in the paradigm of the perception, diagnosis, and treatment of PCOD. The introduction of machine learning applications into the clinical environment is one of the steps toward personalized medicine—the practice that considers both diagnostic and treatment plans based on the physiological characterization, mental condition, and the environment of the individual.

2 Literature Review

The most recent studies emphasize high-precision diagnostic systems using deep learning and ensemble learning techniques. Divekar and Sonawane (2025) developed an AI-enabled ultrasound imaging classifier demonstrating the increasing role of computer vision in gynecological diagnostics [1]. Panjwani et al. (2025) extended this trajectory by proposing optimized ML models incorporating automated feature engineering and hyperparameter tuning for early PCOS detection, achieving improved sensitivity across heterogeneous clinical datasets [2].

Parallel work by Zad et al. (2024) focused on EHR-driven PCOS prediction, leveraging structured medical histories, laboratory findings, and clinical notes to train advanced ML algorithms [3]. Their results validated that patterns embedded in EHR data—such as hormone profiles, metabolic markers, and comorbidities—provide a strong basis for risk prediction. Their approach aligns with the broader trend identified by Wang et al. (2024), who reviewed deep learning models in EHR analytics and emphasized the importance of feature fusion, missing data handling, and temporal modeling in reproductive health prediction tasks [4].

Similarly, Barragán-Montero et al. (2024) presented a cohort-based PCOS prediction study on Chinese clinical data, demonstrating that ML models perform consistently across ethnicities when trained on well-curated datasets [5]. Morris et al. (2023) introduced federated learning as a decentralized approach for PCOS diagnosis, enabling multiple healthcare institutions to collaborate without sharing raw data [6]. Barrera et al. (2023) further reinforced the need for ethical AI in clinical contexts through a systematic review that assessed ML applications in PCOS diagnosis, identifying common limitations such as dataset imbalance, lack of explainability, and inadequate cross-population validation [7].

A significant portion of 2023 literature explored ensemble and stacked learning strategies. Suha and Islam (2023) proposed a modified stacking ensemble that identified dominant PCOS features using feature aggregation and hierarchical classifiers [8]. Joshi et al. (2023) introduced deep linear discriminant analysis with variation for PCOS classification [9]. Hdaib et al. (2023) assessed classical ML algorithms such as SVM, Random Forest, and Decision Trees for PCOS detection [10]. Denny et al. (2023) designed the I-HOPE system, which integrates multiple ML techniques into a unified PCOS prediction dashboard [11]. Alamoudi et al. (2023) advanced the field by presenting a deep-learning fusion model that combines ultrasound images with structured clinical data [12].

Explainable AI (XAI) also gained importance in 2023. Elmannai et al. (2023) developed a feature selection and XAI-based PCOS detection model capable of tracing and justifying diagnostic decisions [13]. Silva et al. (2023) introduced a stacked framework for improved classification accuracy [14]. Ravishankar and Patil (2023) contributed to ovarian cyst detection through a fully convolutional neural network (OCD-FCNN) [15].

The foundational work published in 2022 set the stage for modern ML-driven PCOS research. Nasim et al. (2022) pioneered a bioinformatics-oriented ML pipeline capable of analyzing biochemical and genomics data for PCOS prediction [16]. Danaei Mehr and Polat (2022) employed feature selection techniques to enhance diagnostic precision [17]. Al-

shakrani and Altalhi (2022) proposed a hybrid ML framework incorporating both rule-based and learning-based modules [18]. Tiwari et al. (2022) presented SPOSDS, a smart PCOS diagnostic system using ML [19]. Swamy and Sainadh (2022) developed a hybrid model combining KNN, SVM, and Random Forest [20]. Finally, Suha and Islam (2022) demonstrated the utility of deep learning for ovarian ultrasound image analysis [21].

3 Dataset Description

The research is grounded on a specially assembled dataset of medical and psychological characteristics related to Polycystic Ovarian Disease (PCOD). The dataset was assembled by the combination of questionnaires and certified medical reports of women diagnosed with PCOD and those who were suspected of PCOD but had not been clinically verified. The combination of quantitative medical variables with qualitative self-reported parameters renders the dataset highly appropriate for the development of a machine learning-based diagnostic model.

The dataset constitutes 357 separate records. The dataset consists of around 40 features after data cleaning and feature selection, broadly divided into two categories: Physiological Factors and Psychological Factors.

3.1 Physiological Factors

The physiological sector of the dataset is based on the body symptoms and biological trends that are prevalent in women with PCOD. These involve metabolic, reproductive and functional health indicators. The important physiological characteristics captured in the dataset are summarized in Table 1.

Table 1. Physiological Feature Descriptions

No.	Physiological Factor	Description
1	Daily Activity Performance	Are you able to perform daily activities at a regular pace?
2	Fatigue during Activities	Do you feel fatigue while performing activities?
3	Work Efficiency	Do you experience any effect on your work efficiency?
4	Leisure Activity Participation	Are you able to participate in leisure activities such as walking, jogging, or exercise?
5	Temperature Sensitivity	Is there any impact on your 24-hour temperature perception pattern?
6	Digestion or Bowel Activity	Have you observed any change in your digestion or bowel activity?
7	Pain and Cramps	Do you experience pain and cramps in the lower abdomen or legs?
8	Eating Pattern	Have you observed any change in your eating pattern or appetite?
9	Sleep Quality	Have you experienced any change in the pattern or quality of your sleep?

The answers were gathered on the basis of a Likert scale (1–5, with 1 meaning Never and 5 meaning Always), which allows quantitative analysis. These physiological variables were normalized during preprocessing to reduce the effects of different response scales.

3.2 Psychological Factors

Along with physiological symptoms, the dataset assigns great priority to psychological and emotional data as an increasing number of studies have supported the view that PCOD impacts both physical and mental health. The psychological part of the data was designed to measure these experiences with the help of subjective self-reporting scales. Table 2 describes the psychological characteristics included in the dataset.

Table 2. Psychological Feature Descriptions

No.	Psychological Factor	Description
1	Mood Behavior	Is there any impact of illness on your mood or emotional stability?
2	Memory or Recall	Do you find it difficult to memorize or recall events?
3	Planning Difficulties	Do you face difficulty in planning your daily activities?
4	Decision-Making Difficulty	Do you face difficulty in making decisions?
5	Social Appearance Consciousness	Do you feel conscious of your appearance while interacting socially?
6	Social Withdrawal	Has your social participation decreased after being diagnosed with PCOD?
7	Fertility Stress	Do you feel stressed that you might face trouble getting pregnant?
8	Long-Term Health Concern	Do you feel concerned about the long-term consequences of PCOD?
9	Family Cooperation Fear	Do you fear lack of cooperation from your family members post-PCOD diagnosis?
10	Weight-Related Stress	Do you feel stressed due to your increased weight?

3.3 Dataset Composition and Labeling

The size of the dataset is 357 total records and about 40 attributes. The target variable is dichotomous, derived from two questions: “Do you show signs of PCOD?” and “Do you take any medications against PCOD?” If a participant answered either of these with Yes, Target variable = 1 (PCOD). If both responses were No, the Target variable = 0 (non-PCOD).

3.4 Feature Groups and Distribution

The dataset was split into two major groups of features: a Physiological Group representing all biological and physical health variables, and a Psychological Group reflecting emotional and cognitive health variables. A correlation heatmap demonstrated interesting inter-group relations, indicating that fatigue, abdominal pain, and sleep disturbances were moderately positively related to mood instability, social withdrawal, and fertility-related anxiety.

About 62 percent of the records are of the kind of people diagnosed or suspected of having PCOD, whereas 38 percent were in the control group. This relative imbalance was removed in the process of preprocessing by the use of stratified sampling.

3.5 Statistical Overview

Table 3 provides a brief statistical summary of the dataset before model training.

Table 3. Summary Statistics of Dataset Variables

Category	Count	Mean (Norm.)	Std. Dev.	Missing (%)	Data Type
Physiological Features	20	0.53	0.18	3.2%	Ordinal/Numeric
Psychological Features	20	0.49	0.21	4.1%	Ordinal
Target Variable	1	0.62	0.48	0%	Binary

3.6 Importance of Combined Physiological and Psychological Attributes

The most distinctive feature of the given dataset is that it is dual-domain, combining the physiological and psychological measures in a single diagnostic system. Conventional diagnostic models of PCOD are usually based mainly on the physical symptoms and biochemical parameters. Nevertheless, new studies have revealed that psychological aspects of health, including chronic stress, emotional volatility, and self-image, significantly contribute to disease development and evolution. This unified data model is consistent with the current healthcare trend of bio-psycho-social modeling, in which patient well-being is assessed as a complex of biological, psychological, and social aspects.

4 Methodology

In order to have a reliable and interpretable image related to prediction of Polycystic Ovarian Disease (PCOD), a robust and reproducible machine learning pipeline was created. The methodology framework incorporates stringent data pre-processing, thorough feature selection, and systematization of model training and sound performance validation with multi-split strategies and cross-validation. The main component of the approach is the Gradient Boosting Classifier (GBC), which was selected because of its ability to build nonlinear relationships, address class imbalance, and achieve high predictive efficiencies. The Python 3.10 version with scikit-learn, pandas, and NumPy packages were used to implement each of the phases on a Windows-based computing system.

4.1 Data Acquisition and Feature Identification

The data employed in the current study was constructed on the basis of well-formed questionnaires and verified medical records of 357 women aged 18–40 years. The participants were divided into two major categories: those who had been clinically diagnosed with PCOD and those who had similar symptoms and were not diagnosed. Out of 40 initial variables, 20 most important features were retained according to their diagnostic usefulness, completeness and statistical significance.

4.2 Label Creation

Effective target labeling for a supervised learning model is an essential prerequisite. The target variable (PCOD status) was created by a binary classification process. Two major survey questions were taken as reference: “Do you show signs of PCOD?” and “Do you take medications for PCOD?” If a respondent answered either of these questions with yes, the record was found to be 1 (Positive). If both answers were No, then it was termed 0 (Negative).

4.3 Data Preprocessing

Data preprocessing converts the raw and unstructured data into a format applicable by machine learning algorithms. Because survey-based datasets are likely to be filled with partial or varying responses, pre-processing is essential to hold data consistency and model consistency.

- **Handling Missing Values:** The mode imputation method was used to fill gaps in the data by substituting the most common value in each column. This method works well especially with categorical and ordinal variables, which prevail in the data.
- **Encoding Categorical Variables:** All categorical variables that are textual in nature were encoded into numeric variables through Label Encoding so as to make the dataset machine-readable. This transformation gives an integer to every categorical response (e.g., Yes = 1, No = 0, Sometimes = 2).
- **Feature Consolidation and Normalization:** Min-Max Normalization was used to normalize all features to the same range of 0 to 1, ensuring that features with a larger numeric value do not overwhelm the model. Normalization also increases gradient-based algorithms' convergence speed, which is vital in ensemble models such as Gradient Boosting.
- **Feature Selection:** Correlation coefficients and mutual information scores were used in assessing the importance of features. Multicollinearity was eliminated by removing redundant elements which had correlation coefficients greater than 0.85. The resulting set of 20 high-impact features was determined through this pruning process.

4.4 Model Selection and Training

The selection of machine learning models was informed by the desire to have interpretability, strength, and high accuracy. After comparative experiments with XGBoost, Random Forest, and Support Vector Machines (SVM), the Gradient Boosting Classifier (GBC) was found to be the best fit. Gradient Boosting Classifier repeatedly makes an ensemble of weak learners (decision trees) in which the new model tries to rectify the errors of the previous models. The classifier was applied on scikit-learn using the following hyperparameters:

- `n_estimators = 100`: Specifies the number of boosting steps.
- `learning_rate = 0.1`: Controls the contribution of each tree.
- `max_depth = 3`: Limits the depth of every decision tree to avoid overfitting.
- `random_state = 42`: Enforces inter-run reproducibility.

4.5 Evaluation Strategy

To achieve the objective of determining the robustness of models and avoiding bias during the estimation of performances, several data split strategies were used: 80-20, 66-34, and 50-50. In order to increase the reliability of the evaluation, the training set was subjected to 10-Fold Stratified Cross-Validation. Stratified cross-validation ensures that each fold has a roughly equal number of positive and negative samples, maintaining balance in classes throughout all iterations.

4.6 Performance Metrics

Various measures were used to evaluate model performance: Accuracy (percentage of accurate predictions), Precision (fraction of true positives on all positive predictions), F1-Score (harmonic mean of precision and recall), Cohen's Kappa (inter-rater reliability measure), and Area Under the Curve (AUC). All metrics were computed on each of the three train-test splits and the result averaged across the 10 cross-validation folds.

5 Experimental Setup

Technical implementation of the proposed model, computing environment, and operations performed in order to reach reliable prediction results are described in this section. All experiments were performed in Python (v3.10) in Jupyter Notebook with the aid of libraries like scikit-learn, pandas, NumPy, and matplotlib. The hardware setup consisted of an Intel Core i7 processor with 16GB of RAM and Windows 11 operating system. The Gradient Boosting Classifier was fitted using the following important parameters: Number of Estimators: 100; Learning Rate: 0.1; Max Depth: 3; Cross-Validation Folds: 10; Random State: 42.

5.1 Algorithm Design

The step-by-step process of PCOD prediction based on Gradient Boosting is summarized as follows:

1. **Load Dataset:** Import the dataset using the *pandas* library (`read_excel()` function). Verify column names, check for null values, and inspect data types.
2. **Define Feature Groups:** Two feature lists are defined: Physiological Factors (fatigue, digestion, abdominal pain, sleep irregularity, etc.) and Psychological Factors (stress, fertility anxiety, body image, social withdrawal, etc.).
3. **Filter Columns:** Ensure only valid and available columns are selected from the dataset.
4. **Encode Categorical Variables:** Apply *LabelEncoder* from scikit-learn.
5. **Define Target Variable:** Create binary classification target: if (`Signs_of_PCOD == "Yes"` or `Medication == "Yes"`): Target = 1; else: Target = 0.
6. **Handle Missing Data:** Use `SimpleImputer(strategy='most_frequent')` to replace missing values.
7. **Train-Test Split:** Apply stratified data partitioning using `train_test_split()` with ratios 80-20, 66-34, and 50-50.
8. **Initialize Gradient Boosting Model:** Instantiate `GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)`.
9. **Cross-Validation:** Implement `StratifiedKFold(n_splits=10)`.
10. **Train Final Model:** Fit the model on the full training dataset.
11. **Model Evaluation:** Predict outcomes for the test set and compute all evaluation metrics.
12. **Comparative Benchmarking:** Repeat experiments using alternative classifiers (Logistic Regression, SVM, Random Forest, etc.) to validate the superior performance of Gradient Boosting.

5.2 Implementation Environment

All computations were executed using the **Anaconda distribution** with Jupyter Notebook interface. Libraries used include: pandas for data manipulation and cleaning; numpy for numerical operations; sklearn.preprocessing for encoding and imputation; sklearn.model_selection for data splitting and cross-validation; sklearn.metrics for model evaluation; and matplotlib and seaborn for visualization.

6 Results and Discussion

Findings of this paper provide a comprehensive discussion of the predictive accuracy of several machine learning algorithms in the presence of a structured dataset of physiological and psychological characteristics for the diagnosis of PCOD. The discussion is represented by two main sections: (1) the analysis of the feature importance, based on SHAP (SHapley Additive exPlanations) values to determine the most significant variables in the prediction of PCOD, and (2) comparative analysis of machine learning classifiers.

6.1 SHAP Importance Analysis

The SHAP values measure the contribution that each single feature makes to the final prediction of the model by dividing the marginal contribution of each feature by all the possible combinations of features. The top 20 features contributing to PCOD prediction are summarized in Table 4.

The SHAP analysis highlights several critical insights: (1) The findings indicate that physiological and psychological characteristics play an almost equal role as top predictors, supporting the biopsychosocial model of PCOD; (2) The characteristic "ability to perform daily activities" has the highest SHAP value (0.814567); (3) Fertility anxiety is the second-ranking feature, implying that the psychological state of anxiety regarding fertility may be strongly connected with the underlying hormonal disturbance; (4) Cognitive and emotional disturbances indicate hormonal imbalance; (5) Stress, appearance consciousness, and social withdrawal manifest the socioemotional stigma that women with PCOD commonly experience.

6.2 Model Performance Evaluation

Various classifiers were trained and tested on the dataset under three different train-test splits: 50-50, 66-34, and 80-20. Tables 5, 6, and 7 present the results.

Comparative Discussion Across Splits. In all three settings, the Gradient Boosting Classifier was consistently able to provide the most balanced and accurate predictions. Its ensemble-based structure enabled it to learn fine-grained correlations between psychological and physiological predictors. Random Forest was also a good performer but did not have the sequential correction feature that provides Gradient Boosting its advantage. High precision of SVM suggests good capacity to correctly identify true positives, but recall is comparatively poor. Decision Tree and Naive Bayes models performed poorly due to overfitting sensitivity and feature independence assumptions, respectively.

Metric-Wise Analysis. Gradient Boosting attained accuracy of more than 97% in all splits, showing strong generalization. Gradient Boosting and SVM maintained high precision ($\approx$0.83–0.85), critical in healthcare where false positives can cause unnecessary anxiety. The F1 measure showed the best values in splits for Gradient Boosting, implying effectiveness in classifying true PCOD cases. The Kappa values of 0.52–0.54 confirm that the algorithm is

Table 4. Feature Importance Based on SHAP Values

Rank	Feature Description	Importance Score	Category
1	Are you able to perform daily activities at regular pace as before after being diagnosed with PCOD?	0.814567	Physiological
2	Do you feel stressed that you might face trouble getting pregnant?	0.656562	Psychological
3	Do you find it difficult to memorize or recall events?	0.394938	Psychological
4	Is there any change in your perspective towards your surroundings as before?	0.391881	Psychological
5	Do you experience any effect on your work efficiency?	0.366446	Physiological
6	Do you feel fatigue while performing activities?	0.248033	Physiological
7	Do you feel stressed due to your increased weight?	0.244946	Psychological
8	Do you fear lack of cooperation from your family members post-diagnosis with PCOD?	0.225513	Psychological
9	Is there any impact on your 24-hour pattern of perception of temperature?	0.220027	Physiological
10	Do you feel concerned about the long-term consequences of PCOD?	0.195809	Psychological
11	Are you able to participate in leisure activities as before?	0.187684	Physiological
12	Is there any impact of illness on your mood behaviour?	0.152854	Psychological
13	Do you experience any pain and cramps in your lower abdomen, lower back and legs?	0.142405	Physiological
14	Have you observed any change in your digestion and bowel activity?	0.121643	Physiological
15	Have you experienced any change in the quality and pattern of your sleep?	0.096216	Physiological
16	Do you find any difficulty in planning your daily activities?	0.073863	Psychological
17	Has your social participation decreased after being diagnosed with PCOD?	0.072267	Psychological
18	Do you feel conscious of your appearance while interacting with people socially?	0.069458	Psychological
19	Have you observed any change in your eating pattern?	0.037884	Physiological
20	Do you face any difficulty in making decisions?	0.020542	Psychological

not agreeing by chance. The AUC of Gradient Boosting was 0.76–0.78, the highest among all classifiers.

Holistic Insights and Clinical Implications. The combined effects of the experimental outcomes and SHAP analysis give a detailed view of machine learning to improve the early detection of PCOD. The integration of psychophysiological indicators provides clinicians with an evidence-based rationale to include mental health tests in the PCOD diagnosis. The high predictive accuracy of Gradient Boosting may be used as a back-up diagnostic tool that identifies people at risk prior to formal clinical tests. SHAP explanations enable healthcare professionals to see the causes of every prediction, bridging algorithmic choice and human

Table 5. Model Comparison Using 50-50 Train-Test Split

Model	Accuracy	Precision	F1 Score	Cohen's Kappa	AUC
Gradient Boosting	0.9765	0.8333	0.5556	0.521	0.764
Random Forest	0.9647	0.8333	0.5000	0.476	0.738
SVM	0.9588	0.8571	0.4615	0.451	0.725
Logistic Regression	0.9471	0.8000	0.4444	0.427	0.716
XGBoost	0.9471	0.6667	0.4000	0.401	0.709
KNN	0.9412	0.7500	0.4000	0.388	0.692
Decision Tree	0.9118	0.5000	0.2857	0.308	0.628
Naive Bayes	0.8529	0.5000	0.2105	0.202	0.571

Table 6. Model Comparison Using 66-34 Train-Test Split

Model	Accuracy	Precision	F1 Score	Cohen's Kappa	AUC
Gradient Boosting	0.9706	0.8333	0.5556	0.512	0.765
SVM	0.9647	0.8333	0.5000	0.488	0.750
Random Forest	0.9588	0.8000	0.4706	0.465	0.742
Logistic Regression	0.9471	0.8333	0.4762	0.453	0.738
XGBoost	0.9412	0.7143	0.4000	0.427	0.726
KNN	0.9353	0.6667	0.3636	0.403	0.704
Decision Tree	0.9118	0.6667	0.3077	0.337	0.653
Naive Bayes	0.8824	0.6364	0.2500	0.274	0.604

Table 7. Model Comparison Using 80-20 Train-Test Split

Model	Accuracy	Precision	F1 Score	Cohen's Kappa	AUC
Gradient Boosting	0.9765	0.8462	0.5789	0.541	0.781
SVM	0.9706	0.8462	0.5484	0.522	0.766
Random Forest	0.9647	0.8000	0.5161	0.502	0.752
Logistic Regression	0.9588	0.8333	0.5263	0.496	0.746
XGBoost	0.9529	0.7500	0.4706	0.470	0.733
KNN	0.9412	0.7059	0.4211	0.441	0.719
Decision Tree	0.9118	0.6154	0.3333	0.357	0.666
Naive Bayes	0.8824	0.6667	0.2667	0.304	0.601

judgment. The model's ability to determine which factors have the strongest effect on each specific prediction will allow modifying the treatment plan to be more personal.

7 Conclusion

The Gradient Boosting Classifier emerged as the most reliable, interpretable, and clinically meaningful model for PCOD prediction across all experimental settings. It demonstrated consistent superiority in accuracy, F1, Cohen's Kappa, and AUC metrics, proving its robustness even under varied data partitions. SHAP analysis further validated the model's transparency and provided actionable insights into the psychological and physiological correlates of PCOD.

The exploration of Polycystic Ovarian Disease (PCOD) through an integrated framework of physiological and psychological indicators has yielded valuable insights into the multifaceted nature of the condition and the transformative role of machine learning in women's

healthcare. By unifying medical and behavioral data within a single analytical model, this research moves beyond conventional diagnostic boundaries, emphasizing that PCOD is not solely a hormonal or reproductive disorder, but a biopsychosocial condition that demands holistic evaluation and intervention.

The inclusion of psychological parameters—such as stress, body image perception, anxiety, emotional fatigue, and fertility concerns—proved crucial for achieving high predictive accuracy. The findings reiterate that PCOD is deeply interwoven with psychological wellbeing: women suffering from hormonal imbalance often experience chronic stress, depressive symptoms, and low self-esteem, which can exacerbate metabolic dysfunctions through hormonal feedback mechanisms. This interdependence validates the need for multidimensional assessment protocols that incorporate both physical and emotional indicators when screening for PCOD.

Feature importance analysis indicated that physiological markers such as sleep pattern disturbance, fatigue level, abdominal discomfort, and irregular appetite were among the strongest predictors. However, psychological features—particularly perceived stress, self-image dissatisfaction, and emotional instability—showed nearly equivalent contribution values. This demonstrates that psychological distress does not merely accompany PCOD but can act as a core predictive determinant.

Despite promising results, it is essential to recognize limitations. The dataset, though rich in diversity of variables, involved a relatively modest number of participants (357), which may limit generalization to broader populations. Future research should incorporate larger and longitudinal datasets that track hormonal profiles, mood variations, and lifestyle changes over time. Integrating wearable health monitoring data—such as heart rate variability, sleep tracking, and activity levels—could provide more objective physiological measures that complement psychological inputs.

This study reaffirms that **PCOD is not just a disorder of the ovaries, but of the entire system—endocrine, psychological, and social**. By demonstrating that machine learning can capture this multidimensional complexity with over 90% accuracy, this research bridges the gap between data science and holistic medicine. The future of women's health research lies in embracing this integrative paradigm—where emotional, behavioral, and biological data converge through intelligent algorithms to deliver more humane, precise, and accessible healthcare solutions.

References

[1] S. Divekar, A. Sonawane, Leveraging AI for automatic classification of PCOS using ultrasound imaging. J. Biomed. Eng. AI **17**, 45–53 (2025)

[2] B. Panjwani, J. Yadav, V. Mohan, N. Agarwal, S. Agarwal, Optimized machine learning for the early detection of polycystic ovary syndrome in women. Sensors **25**, 1166 (2025)

[3] Z. Zad, V.S. Jiang, A.T. Wolf, T. Wang, J.J. Cheng, I.C. Paschalidis, S. Mahalingaiah, Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Front. Endocrinol. **15**, 1298628 (2024)

[4] H. Wang, Y. Zhang, L. Liu, Deep learning models for EHR-based predictive tasks: A comprehensive review. arXiv preprint arXiv:2401.00888 (2024)

[5] A. Barragán-Montero, Y. Cheng, J. Li, Machine learning for PCOS prediction using Chinese clinical cohort data. Comput. Biol. Med. **167**, 107527 (2024)

[6] A. Morris, K. Reddy, M. Farooq, Federated learning for privacy-preserving PCOS diagnosis. J. Pers. Med. **13**, 124–132 (2023)

[7] F.J. Barrera, E.D. Brown, A. Rojo, J. Obeso, H. Plata, E.P. Lincango, N. Terry, R. Rodríguez-Gutiérrez, J.E. Hall, S. Shekhar, Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: A systematic review. Front. Endocrinol. **14**, 1106625 (2023)

[8] S.A. Suha, M.N. Islam, Exploring the dominant features and data-driven detection of PCOS through modified stacking ensemble. J. Med. Syst. **47**, 29 (2023)

[9] A. Joshi, R. Gupta, P. Mehta, Deep linear discriminant analysis with variation for PCOS classification. Neural Comput. Appl. (2023). https://doi.org/10.1007/s00521-023-08675-4

[10] M.T. Hdaib, A. Al-Fraihat, M. Al-Smadi, Detection of PCOS using machine learning algorithms. J. Healthc. Eng. **2023**, 1–12 (2023)

[11] P.A. Denny, F. Noor, A. Shahid, I-HOPE: Detection and prediction system for PCOS using machine learning techniques. Procedia Comput. Sci. **218**, 395–402 (2023)

[12] F. Alamoudi, S.A. Khan, R. Alharbi, A deep learning fusion approach to diagnosis of PCOS. Comput. Biol. Med. **152**, 106417 (2023)

[13] W. Elmannai, K. Elleithy, M.S. Alouini, PCOS detection model based on feature selection and XAI. Appl. Soft Comput. **141**, 110262 (2023)

[14] T.A. Silva, F. Sousa, P. Carvalho, Stacked learning framework for accurate classification of PCOS. Biomed. Signal Process. Control **81**, 104534 (2023)

[15] N. Ravishankar, R. Patil, A deep learning approach for ovarian cysts detection and classification (OCD-FCNN). Comput. Biol. Med. **151**, 106354 (2023)

[16] M. Nasim, M.U. Farooq, M.W. Anwar, Novel machine learning approach for PCOS prediction in bioinformatics. Inform. Med. Unlocked **30**, 100950 (2022)

[17] H. Danaei Mehr, K. Polat, Diagnosis of PCOS via feature selection techniques. Expert Syst. Appl. **199**, 117108 (2022). https://doi.org/10.1016/j.eswa.2022.117108

[18] M.S. Alshakrani, A.H. Altalhi, Hybrid machine learning framework for PCOS detection. Diagnostics **12**, 2117 (2022)

[19] S. Tiwari, S. Arora, A. Bajpai, SPOSDS: Smart PCOS diagnostic system using machine learning. Int. J. Med. Inform. **165**, 104855 (2022)

[20] R. Swamy, M.V. Sainadh, Hybrid machine learning for early PCOS prediction. Comput. Electr. Eng. **99**, 107722 (2022)

[21] S.A. Suha, M.N. Islam, An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Sci. Rep. **12**, 16538 (2022)