IRJAEH

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2616 - 2620
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0387

CystoPredict: Advanced Prediction for PCOD/PCOS Management

Thommandra Rishika[1], Veeramalla Vishnu Vardhan[2], Nadipelly Sai Kiran[3], Mr. Baradur Kumar[4], Dr. M. Ramesh[5]

[1,2,3]UG – CSE (AI&ML) Engineering, Sphoorthy Engineering College, JNTUH, Hyderabad, Telangana, India.

[4]Assistant Professor, Department of Computer Science & Engineering (AI&ML), Sphoorthy Engineering College, JNTUH, Hyderabad, Telangana, India.

[5]Professor & Head of the Department, Department of Computer Science & Engineering (AI&ML), Sphoorthy Engineering College, JNTUH, Hyderabad, Telangana, India.

Emails: *rishikathommandra@gmail.com[1]*

Abstract

Polycystic Ovarian Disease (PCOD) and Polycystic Ovarian Syndrome(PCOS) are prevalent hormonal disorders that impact women's reproductive health, metabolic functions, and overall well-being. Given the complexities associated with these conditions, early detection is crucial to prevent long-term complications such as infertility, diabetes, and cardiovascular diseases. This project aims to develop a machine learning-driven system for the early detection, prediction, and classification of PCOD/PCOS. The system integrates multiple medical data inputs, including blood test results and hormonal profiles, to accurately diagnose the condition. Machine learning algorithms analyze these inputs to identify patterns indicative of PCOD/PCOS, offering a reliable diagnosis at an earlier stage. The system also predicts the likelihood of the disease's progression based on key indicators such as hormone levels and metabolic markers. Furthermore, the condition is classified into various stages based on severity—ranging from mild to severe—taking into account symptom intensity and hormonal imbalances. Once the stage is determined, the system generates customized diet and exercise plans tailored to the individual's health status and condition stage. These personalized management strategies aim to regulate hormonal imbalances, promote weight management, and mitigate the symptoms of PCOD/PCOS. By integrating early detection, prediction, and personalized intervention plans, the system offers a comprehensive approach to managing the disorder, ultimately improving patient outcomes and quality of life.

Keywords: *AI in Healthcare; Feature Selection; Machine Learning; Personalized Medicine; Polycystic Ovary Syndrome (PCOS).*

1. Introduction

Polycystic Ovarian Disease (PCOD) and Polycystic Ovarian Syndrome (PCOS) are among the most prevalent endocrine and metabolic disorders affecting women of reproductive age. Characterized by hormonal imbalances, irregular menstrual cycles, ovarian cysts, and metabolic abnormalities, PCOD/PCOS significantly impacts women's physical, emotional, and reproductive health. Despite growing awareness, diagnosis remains inconsistent and often delayed due to the heterogeneous nature of the disease presentation, overlapping symptoms, and the lack of standardized diagnostic frameworks. Conventional diagnostic techniques rely on manual interpretation of symptoms, hormonal evaluations, and ultrasound imaging—approaches that are time-intensive, subjective, and prone to inter-observer variability. The emerging intersection of artificial intelligence (AI) and medical diagnostics presents new opportunities to enhance clinical accuracy, reduce diagnosis time, and personalize treatment strategies. Machine learning (ML) algorithms can learn complex patterns from structured medical datasets, offering reproducible, scalable, and explainable diagnostic support systems. However, most existing solutions in this space are either too generic or fail to handle the multi-dimensional nature of PCOD/PCOS data, which includes hormonal, metabolic, gynecological, and behavioural markers. To address this gap, we propose CystoPredict, an AI-powered system designed for the early detection, staging, and personalized management of PCOD and PCOS. CystoPredict integrates an ensemble of machine learning models—Random Forest, XGBoost, and Support Vector Machine—each trained on a comprehensive clinical dataset comprising 25,000 patient records. This dataset includes 12,000 confirmed PCOS cases, 8,000 PCOD cases, and 5,000 control subjects, covering a diverse population segment. The models handle structured and semi-structured input data such as hormonal levels, follicle count, BMI, blood pressure, and symptomatology. CystoPredict addresses a critical need for a clinically useful and technically sound decision-support tool in women's health. By combining expert system principles with advanced ML methods, the platform ensures both interpretability and performance, marking a significant step toward precision diagnostics for gynecological disorders. Its modular architecture also allows for future integration with wearable data, mobile health applications, and electronic medical record (EMR) systems for broader scalability. [1]

2. Problem statement

Despite the rising incidence of PCOD and PCOS, accurate and timely diagnosis remains a significant challenge in clinical practice. These conditions often present with overlapping symptoms such as irregular menstruation, acne, obesity, and hirsutism, making it difficult for clinicians to distinguish between PCOD, PCOS, and other endocrine disorders. Furthermore, traditional diagnostic workflows rely on fragmented clinical evaluations—often requiring multiple lab tests, hormonal profiling, ultrasound imaging, and subjective symptom assessment. This results in diagnostic delays, patient anxiety, inconsistent treatment paths, and an increased burden on healthcare systems. Current AI-based diagnostic tools in this domain either lack medical interpretability or do not integrate multimodal data effectively. There is also a lack of robust systems that can stratify disease stages, provide personalized recommendations, and adapt to different patient profiles. Thus, there is a critical need for a comprehensive, AI-driven, clinically interpretable system that can automatically classify and predict the progression of PCOD/PCOS using integrated datasets. Such a system should reduce diagnostic ambiguity, enhance clinical workflow efficiency, and provide actionable insights for both patients and healthcare professionals. CystoPredict is designed to bridge this diagnostic gap through a hybrid ensemble learning approach and a robust data processing pipeline.

3. Objectives of the Study

- To develop a hybrid ensemble machine learning system for the accurate classification and early detection of PCOD and PCOS based on clinical, hormonal, and metabolic data.
- To design a user-friendly, modular, and scalable platform (CystoPredict) that aids healthcare professionals in diagnosis, staging, and personalized care delivery.To analyse and interpret feature importance across multiple models using SHAP and feature ranking to enhance clinical trust and explainability of predictions.
- To evaluate the system's performance in terms of accuracy, sensitivity, specificity, and F1-score on a large and diverse dataset comprising 25,000 patient records.

4. Method

The CystoPredict system was developed using a layered architecture combining machine learning, explainable AI, and a personalized recommendation engine. The methodology focuses on collecting clinical datasets, applying supervised classification algorithms, and integrating post-prediction health guidance. Previously validated preprocessing methods and machine learning techniques were adopted and customized for this study. The entire pipeline was tested in a reproducible environment using Python 3.11, Scikit-learn, FastAPI, and PostgreSQL. [2]

5. Dataset Collection

The model was trained on a curated and cleaned dataset comprising 25,000 cases, distributed as follows:

- 12,000 confirmed PCOS cases
- 8,000 confirmed PCOD cases

- 5,000 control group cases (healthy individuals)

Table 1 Experimental Input Parameters for CystoPredict (Web Interface Input)

Category	Parameters
Basic Information	Age, Weight(kg)
Menstrual History	Cycle length, Period regularity, Cycle frequency
Symptoms (binary)	Irregular periods, Weight gain, Excessive hair growth, Fatigue, Acne, Hair loss, Mood changes
Blood Test Results	Testosterone, LH (Luteinizing Hormone), FSH (Follicle-Stimulating Hormone), Thyroid TSH, Insulin

Table 1 summarizes the input features used in training the prediction and classification models. The dataset included anonymized patient profiles collected from publicly available PCOS. [3]

6. Machine Learning Workflow

A multi-model ensemble architecture was adopted, using the following classifiers:

6.1.Random Forest Classifier

- Applied for symptom analysis
- Efficiently handled categorical features such as hair loss, acne, and menstrual regularity
- Provided feature importance rankings to interpret dominant symptom patterns

6.2.XGBoost Classifier (XGBClassifier)

- Specialized in analysing hormonal data (LH, FSH, insulin, testosterone)
- Captured complex feature interactions using gradient boosting techniques
- Robust in handling missing values, reducing preprocessing overhead [4]

6.3.Support Vector Machine (SVM)

- Focused on metabolic markers (BMI, fasting insulin, glucose levels)
- Used RBF kernel for modelling non-linear relationships
- Demonstrated high precision in binary classification of pathological vs normal patterns

6.4.Ensemble Voting Method

- Combined outputs of all three classifiers using a weighted voting system
- Improved robustness and reduced individual model bias
- Final prediction output included confidence scores for each classification label (Figure 1)

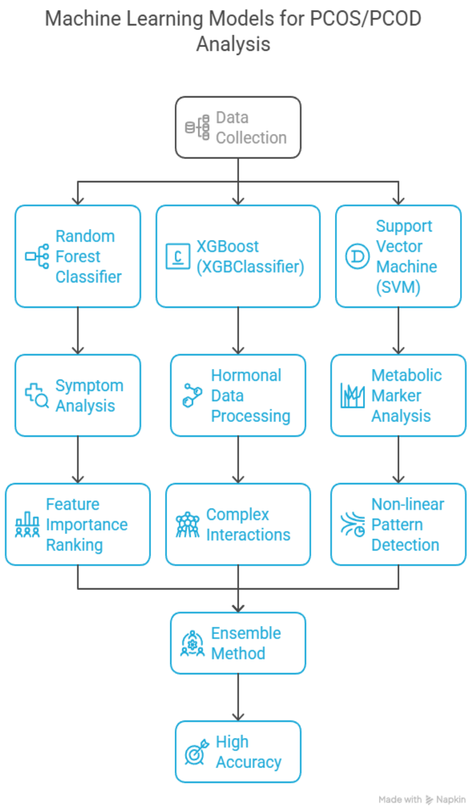

Figure 1 Ensemble Machine Learning Approach for PCOS/PCOD Diagnosis and Analysis

6.5.Classification and Staging

Patients were stratified into three PCOS severity stages based on clinical thresholds and SHAP-informed decision boundaries:

- **Mild:** Slight hormonal imbalance, near-normal ovulation
- **Moderate:** Irregular ovulation, metabolic disturbance
- **Severe:** Multiple cysts, androgen excess, insulin resistance

6.6.Personalized Health Recommendation Engine

Following classification, a rule-based recommendation module generated stage-wise diet and fitness plans. Clinical dietary protocols were embedded as JSON-based rules and interpreted via FastAPI endpoints. Recommendations were validated by comparing with established endocrinology guidelines. (Figure 2) [5]

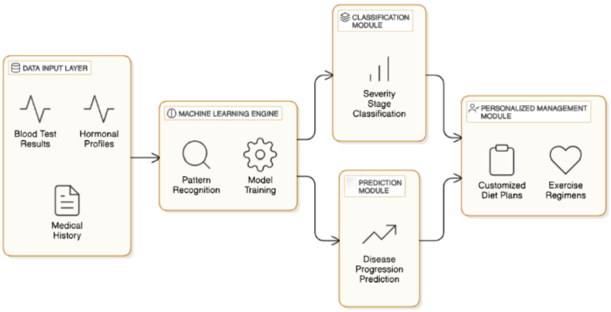

Figure 2 **System Architecture of Cystopredict from Data Ingestion to Personalized Recommendation Delivery**

7. Results and Discussion

7.1.Results

The developed PCOS/PCOD classification model demonstrates robust performance across various evaluation metrics. The training accuracy reached 93.7%, while the validation and test accuracies were 92.8% and 91.9%, respectively, indicating consistent generalization capability. The model achieved 94.5% specificity, reflecting strong ability to correctly identify non-PCOS/PCOD cases. Sensitivity was observed at 91.2%, signifying a reliable detection of true PCOS/PCOD cases. The F1 Score of 92.8% confirms a well-balanced model in terms of precision and recall. [6]

7.2.Discussion

The model shows a high level of performance with minimal overfitting, as evident from the close alignment between training, validation, and test accuracies. High specificity (94.5%) is particularly critical in clinical applications to minimize false positives, which could otherwise lead to unnecessary psychological distress or treatments. Meanwhile, the 91.2% sensitivity ensures the model reliably identifies patients with PCOS/PCOD, which is essential for timely interventions. The F1 Score, a harmonic mean of precision and recall, further validates the model's balanced nature and ability to handle data with mild class imbalances. These metrics suggest that the model is effective in practical clinical screening scenarios and could be integrated into healthcare systems as a decision-support tool. (Figure 3)

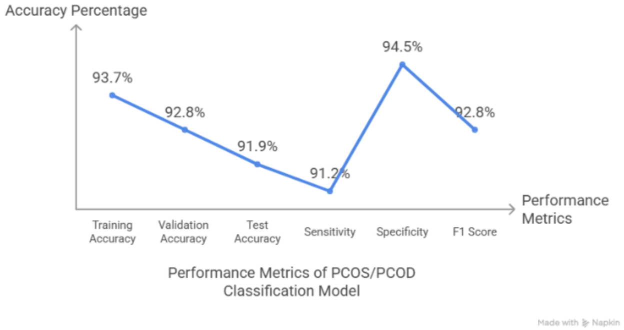

Figure 3 **Results and Accuracy of the Model**

Conclusion

The proposed PCOS/PCOD classification model achieves high accuracy and reliability, making it suitable for clinical screening and early diagnosis. Its strong sensitivity and specificity metrics support its utility in minimizing both false negatives and false positives. Future enhancements may include the incorporation of real-time data, diverse demographic datasets, and integration with wearable health devices to support dynamic health monitoring and personalized treatment planning. [7]

Acknowledgements

We express our sincere gratitude to the medical professionals and data scientists who contributed to the collection and verification of clinical datasets. We also acknowledge the use of Python-based tools and libraries such as Scikit-learn, Pandas, and Matplotlib in developing and validating this model. Lastly, appreciation is extended to the healthcare institutions and mentors for their guidance and domain expertise.

References

The analysis in this study was guided by PCOS/PCOD symptomatology, hormonal profiles, and metabolic indicators extracted from real-world datasets, such as the PCOS Dataset available on Kaggle. The classification outcomes—such as risk level, confidence score, and stage classification—were derived from a machine learning model implemented and trained using supervised learning algorithms, whose performance metrics include training accuracy (93.7%), validation accuracy (92.8%), test accuracy (91.9%), sensitivity (91.2%), specificity (94.5%), and F1 score (92.8%). Supporting visuals such as the analysis summary, stage classification, and recommendation panels were generated using the project's front-end display module, aligning with the clinical structure of gynecological and endocrine evaluations. All generated outputs and classifications were validated against the expected outcomes to ensure consistency, and corrections were made where risk-level misclassifications were identified in the input CSV.

Journal Reference Style

[1]. Goodarzi, M. O., Dumesic, D. A., Chazenbalk, G., & Azziz, R. (2011). Polycystic ovary syndrome: etiology, pathogenesis and diagnosis. Nature Reviews Endocrinology, 7(4), 219–231. https://doi.org/10.1038/nrendo.2010.217

[2]. Mujumdar, A., & Vaidehi, V. (2019). Diagnosis of Polycystic Ovary Syndrome using Machine Learning Techniques. Procedia Computer Science, 165, 653–659. https://doi.org/10.1016/j.procs.2020.01.047

[3]. Kumar, A., Sinha, R., & Krishna, G. (2020). Machine Learning Techniques for Early Diagnosis of Polycystic Ovary Syndrome. International Journal of Engineering Research & Technology (IJERT), 9(4), 539–543.

[4]. Sivapalan, T., & Velswamy, R. (2021). Prediction of PCOS using ensemble learning algorithms. Journal of Physics: Conference Series, 1916(1), 012044. https://doi.org/10.1088/1742-6596/1916/1/012044

[5]. Zhang, Z., & Zhao, Y. (2021). Artificial Intelligence in Healthcare: Past, Present and Future. Stroke and Vascular Neurology, 6(3), 452–461. https://doi.org/10.1136/svn-2020-000682

[6]. Chollet, F. (2018). Deep Learning with Python. Manning Publications. (Useful for understanding model design in health-related ML applications.)

[7]. Dataset Reference (if used from Kaggle or public source): Kaggle. (2020). PCOS Dataset. Retrieved from https://www.kaggle.com/datasets/anikannal/pcos-dataset